IN-64
37985
P-115

NASA Technical Memorandum 104495

A New Numerical Framework for Solving Conservation Laws–The Method of Space-Time Conservation Element and Solution Element

Sin-Chung Chang
National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio

and

Wai-Ming To
Sverdrup Technology, Inc.
Lewis Research Center Group
Brook Park, Ohio

August 1991

NASA

SUMMARY

A new numerical framework for solving conservation laws is being developed. This new approach differs substantially from the well established methods, i.e., finite difference, finite volume, finite element, and spectral methods, in both concept and methodology. It employs:

a. a nontraditional formulation of the conservation laws in which space and time are unified and treated on the same footing; and

b. a nontraditional use of discrete variables such that numerical marching can be carried out by using a set of relations that represents both local and global flux conservation.

To be specific, we consider a conservation law that governs the convection and diffusion of a physical variable in a 1-D space. Let *(i)* x be the spatial coordinate, *(ii)* $c_0 > 0$ be a conversion constant with the dimension of velocity, and *(iii)* t be the product of c_0 and the temporal coordinate. By definition, x and t have the same dimension. As a result, $x_1 = x$ and $x_2 = t$ may be considered as the coordinates of a two-dimensional Euclidean space E_2 (also referred to as a space-time). Let $u(x,t)$ be a scalar function of x and t. Let a be a dimensionless constant and μ (≥ 0) be a constant with the dimension of length. Then the conservation law may be expressed as

$$\oint_{S(V)} \vec{h} \cdot \vec{ds} = 0 \tag{0.1}$$

where *(i)* $S(V)$ is the boundary of an arbitrary volume V in E_2, *(ii)*

$$\vec{h} \stackrel{def}{=} \left(au - \mu \frac{\partial u}{\partial x}, u \right) \tag{0.2}$$

is a current density vector in E_2, and *(iii)* $\vec{ds} = d\sigma \vec{n}$ with $d\sigma$ and $\vec{n}$, respectively, being the area and the outward unit normal of a surface element on $S(V)$. By applying the divergence theorem in E_2, Eq. (0.1) implies the unsteady convection-diffusion equation, i.e.,

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} - \mu \frac{\partial^2 u}{\partial x^2} = 0 \tag{0.3}$$

Let E_2 be divided into nonoverlapping regions referred to as *conservation elements* (see Figs. 2.1(a) and 2.1(b)). A conservation element and its interior, respectively, may be denoted by CE(j,n) and CE$''(j,n)$ where j and n, respectively, are the spatial and temporal indices. For $(x,t) \in$ CE$''(j,n)$, $u(x,t)$ will be approximated by

$$\underline{u}(x,t) \stackrel{def}{=} \alpha_j^n (x - x_j^n) + \beta_j^n (t - t^n) + \gamma_j^n \tag{0.4}$$

where α_j^n, β_j^n and γ_j^n are constants in CE$''(j,n)$, and (x_j^n, t^n) are the coordinates of the center of

CE(j,n). For $(x,t) \in$ CE$''(j,n)$, $h(x,t)$ will be approximated by

$$\vec{\underline{h}}(x,t) \stackrel{def}{=} \left(a\underline{u}(x,t) - \mu \frac{\partial \underline{u}(x,t)}{\partial x} , \underline{u}(x,t) \right) \tag{0.5}$$

Furthermore, the conservation law Eq. (0.1) will be approximated by

$$\oint_{S(\underline{V})} \vec{\underline{h}} \cdot \vec{ds} = 0 \tag{0.6}$$

where $\underline{V}$ is the union of any combination of conservation elements. Since $\vec{\underline{h}}$ is not defined on $S(\underline{V})$, the above surface integration, by definition, is to be carried out over a surface which is in the interior of $\underline{V}$ and immediately adjacent to $S(\underline{V})$.

Because $\underline{u}(x,t)$ and $\vec{\underline{h}}(x,t)$ are continuous in the interior of a conservation element but may be discontinuous across an interface separating two neighboring conservation elements, a conservation element is also a *solution element* in the current scheme. As will be shown, generally a conservation element is not necessarily a solution element and vice versa.

Let $\underline{V}$ = CE(j,n). Then Eqs. (0.4) - (0.6) and the divergence theorem imply that $\beta_j^n = -a\,\alpha_j^n$. As a result, Eq. (0.4) can be simplified as

$$\underline{u}(x,t) = \alpha_j^n [(x - x_j^n) - a(t - t^n)] + \gamma_j^n \quad , \qquad (x,t) \in \text{CE}''(j,n) \tag{0.7}$$

Thus, for $(x,t) \in$ CE$''(j,n)$, $\underline{u}(x,t)$ is determined by the parameters γ_j^n and α_j^n. As will be shown, by repeatedly applying Eq. (0.6) with $\underline{V}$ being the union of two neighboring conservation elements, any pair of γ_j^n and α_j^n can be determined in terms of γ_j^0 and α_j^0, $j = 0, \pm 1, \pm 2, \cdots$. The values of γ_j^0 and α_j^0, in turn, can be determined by the initial condition.

Because

$$\underline{u}(x_j^n, t^n) = \gamma_j^n \tag{0.8}$$

and

$$\frac{\partial \underline{u}}{\partial x} = \alpha_j^n \qquad (x,t) \in \text{CE}''(j,n) \tag{0.9}$$

γ_j^n and α_j^n, respectively, may be considered as the numerical counterparts of $u(x_j^n, t^n)$ and $u_x(x_j^n, t^n)$. In other words, both u and *its spatial derivative* at (x_j^n, t^n) are computed by the current scheme.

In the current paper, we also

a. explore the concept of a dynamic space-time mesh (the conservation elements are embedded in this mesh) and the need for a unified treatment of physical variables and mesh parameters;

b. study the stability, dissipation and dispersion of the current scheme by using a rigorous Fourier analysis;

c. develop a new error analysis technique that enables us to predict and interpret the numerical errors of the current and other classical schemes;

d. study the consistency and truncation error of the current scheme; and

e. compare the errors of the numerical solutions generated by the current scheme and other classical schemes.

The key results obtained from the above study are:

a. It is demonstrated that *(i)* stability and accuracy can be improved, and *(ii)* dissipation and dispersion can be reduced, if the space-time mesh is allowed to evolve with the physical variable such that the local convective motion of physical variables relative to the moving mesh is kept to a minimum. *Because the appearance of wiggles near a discontinuity is a result of numerical dispersion, these wiggles can also be reduced by reducing this relative convective motion.*

b. It is shown that *there is a remarkable similarity between the forms of the amplification factors of the Leapfrog/DuFort-Frankel and the current schemes.* As a result of this similarity, the stability condition of the current scheme, as in the case of the Leapfrog/DuFort-Frankel scheme, *is essentially the CFL condition and thus independent of the viscosity* μ. Therefore, the current scheme *is unconditionally stable in the case of pure diffusion.* Also, as in the case of the Leapfrog/DuFort-Frankel scheme, the current scheme *has no numerical diffusion in the absence of viscosity.* Note that the stability condition of a classical explicit scheme for solving Eq. (0.3), e.g., the MacCormack scheme, generally is more restrictive than the CFL condition. In the case where the mesh Reynolds number $\ll 1$, the stability bound for the time-step size Δt is more or less proportional to $(\Delta x)^2$. In contrast, the same bound will still be determined by the CFL condition and therefore is proportional to Δx if the current scheme is used. *The advantage of the current scheme in the allowable time-step size grows as* $\Delta x \rightarrow 0$. *This advantage becomes particularly important when the current scheme is used in a steady-state calculation.*

c. It is shown theoretically that the current scheme is more accurate than the Leapfrog/DuFort-Frankel scheme by one order (in a sense to be defined in the paper) in both initial-value specification and the marching scheme. It is also shown theoretically that the current scheme is substantially more accurate than the MacCormack scheme in spite of their almost identical operation counts. Its advantage ranges from a factor of four for the case of pure convection to several orders of magnitude if diffusion is dominant and a theoretically-determined optimal Δt is used.

d. It is shown that the consistency of the current scheme, as in the case of the Leapfrog/DuFort-Frankel scheme, requires that $\Delta t/\Delta x \to 0$ as $\Delta t, \Delta x \to 0$. This fact contrasts sharply with most other explicit schemes, e.g., the MacCormack scheme, that have no such requirement for consistency. However, by using Lax's equivalence theorem and a necessary condition for convergence, it is shown that, for these explicit schemes, *this requirement must manifest itself as a part of stability conditions.* As a matter of fact, it is shown that *the truncation errors of the Leapfrog/DuFort-Frankel, the MacCormack, and the current schemes are all second order in Δx if stability is taken into consideration.*

e. It is shown numerically that the current scheme is far superior to the Leapfrog/DuFort-Frankel scheme in accuracy, and has a substantial advantage over the MacCormack scheme in both accuracy and stability.

1. INTRODUCTION

This paper is the first of a series of papers that describe a new framework for solving conservation laws. This new approach differs substantially from the well established methods, i.e., finite difference, finite volume, finite element, and spectral methods, in both concept and methodology. The focus of the current numerical simulation is entirely on the integral forms of the conservation laws. Little or no attempt is made to simulate the differential forms which are valid only when the dynamical variables are well behaved. As a result, this new framework has the potential to provide more accurate simulation of the physical phenomena in which the dynamical variables may not vary smoothly.

Specifically, the explicit scheme to be presented in this paper employs:

a. a nontraditional formulation of the conservation laws in which space and time are unified and treated on the same footing; and

b. a nontraditional use of discrete variables such that numerical marching can be carried out by using a set of relations that represents both local and global flux conservation.

As a preliminary, this paper will begin with a discussion on the conservation laws. For simplicity, we consider the conservation law that governs the convection of a physical variable in a 1-D space. Let *(i)* x be the spatial coordinate, *(ii)* $c_0 > 0$ be a conversion constant with the dimension of velocity, and *(iii)* t be the product of c_0 and the temporal coordinate. By definition, x and t have the same dimension. As a result, $x_1 = x$ and $x_2 = t$ may be considered as the coordinates of a two-dimensional Euclidean space E_2 [p.161, 1]. Note that the scalar product of any two vectors in E_2 is defined in [1] as a part of the definition of E_2. Let $u(x,t)$ be a scalar function of x and t. Let a be a dimensionless constant. Then the conservation law can be expressed as

$$\oint_{S(V)} \vec{h} \cdot \vec{ds} = 0 \tag{1.1}$$

where *(i)* $S(V)$ is the boundary of an arbitrary volume V in E_2, *(ii)*

$$\vec{h} \stackrel{def}{=} (au, u) \tag{1.2}$$

is a current density vector in E_2, and *(iii)* $\vec{ds} = d\sigma \vec{n}$ with $d\sigma$ and $\vec{n}$, respectively, being the area and the outward unit normal of a surface element on $S(V)$. Note that an n-dimensional Euclidean space E_n ($n \geq 2$) may be referred to as a space-time if one of its coordinates is temporal in nature while others are spatial in nature. Also $\vec{h} \cdot \vec{ds}$ may be referred to as the flux of $\vec{h}$ leaving the volume V through the surface element $\vec{ds}$. With the aid of the divergence theorem and the fact that $\vec{\nabla} \cdot \vec{h} = \partial(au)/\partial x + \partial u/\partial t$, one may obtain the differential form of Eq. (1.1), i.e.,

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \tag{1.3}$$

The current unified and equal treatment of space and time is in sharp contrast to the traditional approach in which space and time are divided and treated separately. The following comments are made to clarify the differences between these two approaches:

a. Geometric figures referred to in the traditional approach, such as rectangles and triangles, usually are objects in space (Note: By definition, "space" is the coordinate hyperplane with time = 0. See p.263 in [2]). Contrarily, geometric figures referred to in the current paper, unless specified otherwise, are objects in space-time. Note that, in this paper, a geometric figure, such as a rectangle, implies both its boundary and interior.

b. In a space-time E_2, the volume V in Eq. (1.1) is traditionally taken to be a rectangle with its edges being aligned with either the $x-$ axis or the $t-$ axis. With this choice, the integral on the left side of Eq. (1.1) can be divided into four parts, each of which involves only integration in time or space. Contrarily, in the current approach, the volume V can be a geometric figure of any shape and thus the surface integration over the boundary of V may involve both space and time simultaneously.

c. In a space-time E_n with $n \geq 3$, we also can consider a conservation law with the form of Eq. (1.1). For example, the mass conservation law in a space-time E_3 can be expressed in the form of Eq. (1.1) with the coordinates x_1, x_2 and x_3, and the vector $\vec{h}$ defined by

$$x_1 \stackrel{def}{=} x \ , \qquad x_2 \stackrel{def}{=} y \ , \qquad x_3 \stackrel{def}{=} t \tag{1.4}$$

and

$$\vec{h} = \left(\frac{v_x}{c_0}\rho \ , \ \frac{v_y}{c_0}\rho \ , \ \rho \right) \tag{1.5}$$

Here *(i)* y is a spatial coordinate, *(ii)* ρ is the mass density, and *(iii)* v_x and v_y are the velocity components in the $x-$ and $y-$ directions, respectively. In the traditional approach, the volume V in Eq. (1.1) is taken to be a cylinder in E_3 like that depicted in Fig. 1.1. Assuming that *(i)* each of the two ends of this cylinder has a constant value of t, and *(ii)* the generators of its side surface point in the $t-$ direction, then Eq. (1.1) implies that

$$\left[\oint_{V_\perp} \rho \, dv_\perp \right]_{(t+\Delta t)} - \left[\oint_{V_\perp} \rho \, dv_\perp \right]_t + \frac{1}{c_0} \int_t^{t+\Delta t} dt' \int_{S_\perp(V_\perp)} \rho \vec{v} \cdot \vec{ds}_\perp = 0 \tag{1.6}$$

where *(i)* $V_\perp$ is the projection of the cylinder on the x–y plane, *(ii)* $dv_\perp$ is a volume element in $V_\perp$, *(iii)* $S_\perp(V_\perp)$ is the boundary of $V_\perp$, *(iv)* $\vec{ds}_\perp$ is a surface element on $S_\perp(V_\perp)$, and *(v)* $\vec{v} \stackrel{def}{=} (v_x, v_y, 0)$ is a vector lying on the x–y plane. Traditionally, $V_\perp$ is

referred to as the "control volume" and Eq. (1.6) is known as the integral form of the two-dimensional time-dependent mass conservation law. Note that, by definition, t/c_0 is the ordinary temporal coordinate. Since the control volume $V_\perp$ is an object in space, the first two integrals in the above equation involve only integration in space while the last integral represents a combined operation in which a surface integration in space is followed by an integration in time. This is another example in which space and time are divided and treated separately. With the division, the above equation has a simple interpretation, i.e., the increase of the mass in the control volume $V_\perp$ during a time interval $\Delta t/c_0$ is equal to the total mass entering $V_\perp$ through its boundary $S_\perp(V_\perp)$ during the same time interval. However, the division of space and time is achieved at the expense of limiting the choice of the volume V to a cylinder in space-time like that depicted in Fig. 1.1. Note that $\rho\vec{v}$, a vector in space, is commonly known as the mass current density vector. Also $\rho\vec{v}\cdot\vec{ds}_\perp$ is referred to as the flux of $\rho\vec{v}$ leaving the control volume $V_\perp$ through the surface element $\vec{ds}_\perp$. These definitions involve only vectors in space. On the contrary, the current density vector $\vec{h}$ is a vector in *space-time* while the flux $\vec{h}\cdot\vec{ds}$ is the inner product of two vectors in *space-time.*

The above remarks make it clear that a greater flexibility in choosing the volume V is allowed in the current formulation of conservation laws than that allowed in the traditional approach. As will be shown, the use of this flexibility is an integral part of the current numerical framework.

Next the classical Lax-Wendroff scheme will be discussed using the uniform mesh depicted in Fig. 1.2(a). This discussion is presented such that readers may understand the stream of thoughts that leads to the development of the current framework.

The Lax-Wendroff scheme for solving Eqs. (1.1) and (1.2) consists of two distinctly different marching steps. In the first step, the variables $u_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ at the time level $(n+\frac{1}{2})$ are evaluated in terms of the variables u_j^n at the time level n, i.e.,

$$u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{2}\left[(1+\nu)u_j^n + (1-\nu)u_{j+1}^n\right] \tag{1.7}$$

where

$$\nu \stackrel{def}{=} a\Delta t/\Delta x \tag{1.8}$$

is the Courant number. The derivation of Eq. (1.7) may be explained using Fig. 1.2(a). Point Q is at the time level n and on the same characteristic line with the mesh point P. The value of u at P, i.e., $u_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ is evaluated in terms of those at the mesh points R and S, i.e., u_{j+1}^n and u_j^n, by assuming *(i)* u is constant along a characteristic line, and *(ii)* the value of u at Q is the linear interpolation of those at R and S.

In the second step, the variables u_j^{n+1} at time level $n+1$ are determined in terms of the variables at time levels n and $n+\frac{1}{2}$ by using a conservation relation. Specifically, one assumes that u_j^n, $u_{j+\frac{1}{2}}^{n+\frac{1}{2}}$, u_j^{n+1} and $u_{j-\frac{1}{2}}^{n+\frac{1}{2}}$, respectively, represent the average values of u on the line segments $\overline{BC}$, $\overline{CD}$, $\overline{DA}$ and $\overline{AB}$. Then an application of Eq. (1.1) with V being the rectangle ABCD implies that

$$u_j^{n+1}\,\Delta x - u_j^n\,\Delta x + a\,u_{j+\frac{1}{2}}^{n+\frac{1}{2}}\,\Delta t - a\,u_{j-\frac{1}{2}}^{n+\frac{1}{2}}\,\Delta t = 0 \tag{1.9}$$

Eq. (1.9) is equivalent to

$$u_j^{n+1} = u_j^n - \nu\left(u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - u_{j-\frac{1}{2}}^{n+\frac{1}{2}}\right) \tag{1.10}$$

This is the relation used in the second marching step. Substituting Eq. (1.7) into Eq. (1.10), one obtains a difference form in which the variables at time level $n+1$ are expressed directly in terms of the variables at time level n, i.e.,

$$u_j^{n+1} = \frac{\nu(\nu+1)}{2}\,u_{j-1}^n + (1-\nu^2)\,u_j^n + \frac{\nu(\nu-1)}{2}\,u_{j+1}^n \tag{1.11}$$

To serve as the starting point of the current development, the conservation relation (1.9) will be cast into a form similar to Eq. (1.1). To proceed, let (see Fig. 1.2(b))

$$\underline{u}(x,t) \overset{def}{=} \begin{cases} u_j^n & \text{if } (x,t) \in \Diamond'' \text{IBGC} \\ u_{j+\frac{1}{2}}^{n+\frac{1}{2}} & \text{if } (x,t) \in \Diamond'' \text{DICF} \end{cases} \tag{1.12}$$

where $\Diamond''$ IBGC denotes the interior of the rhombus $\Diamond$ IBGC, and so on. Similarly, $\underline{u}(x,t)$ can be defined for any (x,t) which is in the interior of other similar rhombuses. Since $\underline{u}(x,t)$ is continuous in the interior of each of these rhombuses but may be discontinuous across an interface separating two neighboring rhombuses, such a rhombus will be referred to as a *solution element*. In terms of $\underline{u}(x,t)$, the vector function $\vec{\underline{h}}(x,t)$ is defined by

$$\vec{\underline{h}}(x,t) \overset{def}{=} \left(a\underline{u}(x,t)\,,\,\underline{u}(x,t)\right) \tag{1.13}$$

With the aid of Eqs. (1.12) and (1.13), Eq. (1.9) can be expressed as

$$\oint_{S(\square ABCD)} \vec{\underline{h}} \cdot \vec{ds} = 0 \tag{1.14}$$

i.e., the total flux leaving the rectangle $\square$ ABCD vanishes if $\vec{\underline{h}}$ is the flux density vector. Note that $\vec{\underline{h}}$ is not defined at the vertices A, B, C and D. However, contributions to the above integral from these isolated points are zero no matter what $\vec{\underline{h}}$ are assigned to them. As a result, they may simply be excluded from the above surface integration. Since Eq. (1.9) applies for any $j = 0, \pm1, \pm2, \cdots$ and $n = 0, 1, 2, \cdots$, Eq. (1.14) is valid if $\square$ ABCD is replaced by any similar rectangle like $\square$ JADK or $\square$ DCML shown in Fig. 1.2(b). For this reason, each of these rectangles will be referred to as a *conservation element* for the Lax-Wendroff scheme. Note that, excluding

its two end points, an interface separating two neighboring conservation elements is located in the interior of a solution element. As a result, the vector function $\vec{\underline{h}}$ is continuous across such an interface. This coupled with Eq. (1.14) implies that the total flux leaving any volume $\underline{V}$ which is the union of any combination of conservation elements must also vanish, i.e.,

$$\oint_{S(\underline{V})} \vec{\underline{h}} \cdot d\vec{s} = 0 \tag{1.15}$$

For example, $\underline{V}$ can be the L-shaped figure formed by □ABCD, □JADK and □DCML. Eq. (1.15) is in a form similar to Eq. (1.1). However, it is equivalent to Eq. (1.9) and thus represents *only one of two marching steps that form the Lax-Wendroff scheme.*

At this juncture, we emphasize that both solution elements and conservation elements are domains in space-time. Contrarily, elements in the finite element method are domains in space only.

In its earliest form, the current scheme may be considered as a modification of the Lax-Wendroff scheme in which *all the marching steps are derived from a single conservation relation.* The modifications begin with the assumption that $u(x,t)$ is approximated by

$$\underline{u}(x,t) = \alpha_j^n (x - x_j^n) + \beta_j^n (t - t^n) + \gamma_j^n \qquad \text{if } (x,t) \in \Diamond'' \text{ IBGC} \tag{1.16}$$

where *(i)* (x_j^n, t^n) are the coordinates of the center of $\Diamond$ IBGC depicted in Fig. 1.2(b), and *(ii)* α_j^n, β_j^n and γ_j^n are considered constants in $\Diamond''$ IBGC. Note that x_j^n is only a function of j if a "stationary" space-time mesh, e.g., a mesh shown in Fig. 1.2(b), is considered. However, in Section 2, it becomes a function of both j and n when a "moving" mesh is introduced. Also note that

$$\underline{u}(x_j^n, t^n) = \gamma_j^n \tag{1.17a}$$

and

$$\frac{\partial \underline{u}}{\partial x} = \alpha_j^n \quad \text{and} \quad \frac{\partial \underline{u}}{\partial t} = \beta_j^n \qquad \text{if } (x,t) \in \Diamond'' \text{ IBGC} \tag{1.17b}$$

i.e., γ_j^n is the value of $\underline{u}$ at the center of $\Diamond$ IBGC while α_j^n and β_j^n, respectively, are the spatial and temporal derivatives of $\underline{u}$ in $\Diamond''$ IBGC.

Hereafter, unless specified otherwise, an equation like Eq. (1.16) is assumed to be valid for any (j,n) with either *(i)* $n = 0, 1, 2, \cdots$, $j = 0, \pm 1, \pm 2, \cdots$, or *(ii)* $n = 1/2, 3/2, 5/2, \cdots$, $j = \pm 1/2, \pm 3/2, \pm 5/2, \cdots$. Thus the rhombuses referred to earlier are also the solution elements in the current method.

In the current method, it is also assumed that

$$\oint_{S(\underline{V})} \vec{\underline{h}} \cdot \vec{ds} = 0 \tag{1.18}$$

where

$$\vec{\underline{h}}(x,t) \stackrel{def}{=} (a\underline{u}(x,t) , \underline{u}(x,t)) \tag{1.19}$$

is the flux density vector, and $\underline{V}$ is the union of *any combination of the rhombuses referred to earlier.* Since $\vec{\underline{h}}$ is not defined on $S(\underline{V})$, the above surface integration, by definition, is to be carried out over a surface which is in the interior of $\underline{V}$ and immediately adjacent to $S(\underline{V})$. A necessary condition for the conservation relation Eq. (1.18) is

$$\oint_{S(\Diamond)} \vec{\underline{h}} \cdot \vec{ds} = 0 \tag{1.20}$$

where $\Diamond$ is any one of the rhombuses referred to earlier. *Thus these rhombuses are also the conservation elements in the current scheme. This is different from the Lax-Wendroff scheme in which a conservation element is a rectangle like* □ABCD *depicted in Fig. 1.2(b).*

Another necessary condition for Eq. (1.18) is the requirement that the net flux of $\vec{\underline{h}}$ entering an interface separating two neighboring conservation elements (i.e., the rhombuses) must vanish. This may be seen by applying Eq. (1.18) separately to two neighboring rhombuses and then to the union of them. Obviously the local flux conservation relations at each interface and within each conservation element (i.e., Eq. (1.20)) are equivalent to the conservation relation Eq. (1.18). In the next section, the current marching procedure will be constructed by using the local conservation relations.

This completes the description of the basic concepts behind the current development. In this first paper, these concepts will be used to construct a numerical scheme for solving an unsteady 1-D constant-coefficient convection-diffusion model equation over a uniform constant-velocity moving mesh. The model equation and the mesh used are simple enough such that the important properties of the resulting scheme may be studied analytically. Yet they are complicated enough that the information gained and the techniques developed in the current study may provide a solid base for the development of new schemes for solving nonlinear conservation laws in higher dimension. Note that it has been shown empirically that the local behaviors of a nonlinear variable-mesh scheme may be studied by using a local analysis (such as the von Neumann analysis) in which the dynamic coefficients and geometric parameters are frozen at their local values. In the same spirit, *the current analysis is intended to serve as a guide for the local analysis of the more complicated schemes to be developed later.*

The remainder of this paper is briefly described as follows: In Section 2, we construct the current scheme without using several questionable assumptions commonly made in the construction of an explicit, time-accurate, conservative scheme. We also point out several fundamental differences that separate the current scheme from the traditional schemes. One of

them is the fact that the current scheme is locally implicit while globally explicit. It is also explained how these differences will result in greater stability and accuracy for the current scheme.

In Section 3, we explore the concept of a dynamic space-time mesh and the need for a unified treatment of the physical variables and mesh parameters. Specifically, it is demonstrated that the stability and accuracy of a numerical calculation may be improved if the space-time mesh is allowed to evolve with the physical variables such that the local convective motion of the physical variables relative to the moving mesh is kept to a minimum. In the meantime, a parameter defined in Section 2 is interpreted as the Courant number for a moving mesh.

In Section 4, the stability, dissipation and dispersion of the current scheme are studied using a rigorous Fourier analysis. It is shown that there is a remarkable similarity between the forms of the amplification factors of the Leapfrog/DuFort-Frankel [p.161, 3] and the current schemes. Note that, hereafter, the former will be referred to as the L/D-F scheme. As a result of this similarity, the stability condition of the current scheme, as in the case of the L/D-F scheme, *is essentially the CFL condition and thus independent of the viscosity coefficient* μ. Therefore, the current scheme *is unconditionally stable in the case of pure diffusion.* Also, as in the case of the L/D-F scheme, the current scheme *has no numerical diffusion in the absence of viscosity.* Note that the stability condition of a classical explicit scheme for solving Eq. (2.2), e.g., the MacCormack scheme [p.163, 3], generally is more restrictive than the CFL condition (see Fig. 4.1). In the case that the mesh Reynolds number $\ll 1$, the stability bound for Δt is more or less proportional to $(\Delta x)^2$. In contrast, the same bound will still be determined by the CFL condition and therefore is proportional to Δx if the current scheme is used. *The advantage of the current scheme in the allowable time-step size grows as* $\Delta x \rightarrow 0$. *This may become particularly important when the current scheme is used in a steady-state calculation.*

In Section 5, assuming smooth and periodic initial data, an error analysis technique is developed using the discrete Fourier analysis formulated in Section 4. The main achievement in this development is the derivation of a simple formula for predicting the numerical errors of the current scheme. This formula contains a principal part and a spurious part. The principal part grows linearly with the time-step number n while the spurious part is independent of n. Thus the principal part will become dominant as n increases. Furthermore, it will be shown that this error-prediction formula is valid up to any n as long as the numerical solution is still accurate up to this n. Similar error-prediction formulae are also given for the L/D-F and the MacCormack schemes. The prediction formula for the L/D-F scheme also contains a principal part and a spurious part while that for the MacCormack scheme contains only the principal part. By using these formulae, it will be shown that the current scheme is more accurate than the L/D-F scheme by one order (in a sense to be defined later) in both initial-value specification and the marching scheme. These formulae may also be used to show that the current scheme is substantially more

accurate than the MacCormack scheme. This section is concluded by showing that the operation counts for the current scheme and the MacCormack scheme are almost identical.

In Section 6, it is shown that the consistency of the current scheme, as in the case of the L/D-F scheme, requires that $\Delta t/\Delta x \to 0$ as $\Delta t, \Delta x \to 0$. This contrasts sharply with most other explicit schemes, e.g., the MacCormack scheme, which have no such requirement for consistency. However, by using Lax's equivalence theorem [p.45, 4] and a necessary condition for convergence, it is shown that, for these explicit schemes, this requirement must manifest itself as a part of the stability conditions. As a matter of fact, it is shown that the truncation errors of the MacCormack, the L/D-F, and the current schemes are all second order in Δx if stability is taken into consideration.

In Section 7, numerical solutions generated by the MacCormack, the L/D-F, and the current schemes are compared with the corresponding analytical solutions for different values of physical coefficients, mesh parameters and total running time. These comparisons show that the current scheme is far superior than the L/D-F scheme in accuracy, and has a substantial advantage over the MacCormack scheme in both accuracy and stability. Moreover, they confirm many of the theoretical predictions made earlier in this paper.

Finally, odds and ends are dealt with in Section 8. They include discussions on boundary-value specification, conservation elements of other geometric shapes, and a possible extension of the current scheme to a space-time of higher dimension.

2. MARCHING PROCEDURES

In Section 1, for simiplicity, we consider only pure convection. Thus the flux density vector $\vec{h} = (au, u)$. Hereafter both convection and diffusion will be considered. As a result, Eq. (1.2) is replaced by

$$\vec{h} \stackrel{def}{=} \left(au - \mu \frac{\partial u}{\partial x}, u\right) \tag{2.1}$$

where $\mu \geq 0$ is a constant with the dimension of length. We will continue to assume Eq. (1.1) and the related assumptions. Note that the unsteady convection-diffusion equation

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} - \mu \frac{\partial^2 u}{\partial x^2} = 0 \tag{2.2}$$

follows from Eq. (1.1) and (2.1) if u is well-behaved.

For reasons to be explained later, we will also consider a moving mesh shown in Fig. 2.1(a). By "moving mesh", we mean a space-time mesh such that the coordinate x may vary along a j mesh line, i.e., a mesh line with a constant value of the index j. In Fig. 2.1(a), b is a constant and $dx/dt = b$ along any j mesh line. In other words, a particle with a space-time trajectory coinciding with a j mesh line has a constant velocity b. For this reason, b may be referred to as the velocity of the moving mesh. The moving mesh is reduced to a stationary mesh if $b = 0$. Let the origin of the coordinate system coincide with the mesh point with $j = n = 0$. Then the coordinates x and t for a mesh point (j,n) are given by

$$x = x_j^n \stackrel{def}{=} j\Delta x + n\, b\Delta t \qquad \text{and} \qquad t = t^n \stackrel{def}{=} n\Delta t \tag{2.3}$$

In the current method, a conservation element is a parallelogram like that depicted in Fig. 2.1(b). It is also a solution element. Hereafter, no distinction will be made between a conservation element and a solution element. A conservation element which is centered at (x_j^n, t^n) will be denoted by CE(j,n). Its interior will be denoted by CE$''(j,n)$.

To construct the marching procedure, $\underline{u}(x,t)$ is assumed to be in the form defined by Eq. (1.16) for $(x,t) \in$ CE$''(j,n)$. Moreover, we assume the conservation relation Eq. (1.18) with

$$\vec{\underline{h}}(x,t) \stackrel{def}{=} \left(a\underline{u}(x,t) - \mu \frac{\partial \underline{u}(x,t)}{\partial x}, \underline{u}(x,t)\right) \tag{2.4}$$

where $\underline{V}$ is the union of any combination of conservation elements. Obviously, this assumption is again equivalent to Eq. (1.20) and the interface flux conservation condition referred to in Section 1. The only modification required is the generalization of conservation elements from rhombuses to parallelograms.

With the above assumptions, one has

$$\vec{\nabla}\cdot\vec{\underline{h}} \stackrel{def}{=} \frac{\partial}{\partial x}\left(a\underline{u}-\mu\frac{\partial \underline{u}}{\partial x}\right)+\frac{\partial \underline{u}}{\partial t} = a\,\alpha_j^n + \beta_j^n \tag{2.5}$$

Thus the diffusion term in $\vec{\underline{h}}$ does not contribute to $\vec{\nabla}\cdot\vec{\underline{h}}$. However, the diffusion term does play a role in the flux balance across an interface. With the aid of Eq. (2.5) and the divergence theorem, the generalized form of Eq. (1.20) implies that $a\alpha_j^n + \beta_j^n = 0$. As a result, Eq. (1.16) implies that

$$\underline{u}(x,t) = \alpha_j^n\left[(x-x_j^n)-a(t-t^n)\right]+\gamma_j^n \qquad \text{if} \quad (x,t)\in \mathrm{CE}''(j,n) \tag{2.6}$$

As a preliminary to the application of the interface flux conservation relation, next we consider the problem of evaluating the flux $\vec{\underline{h}}\cdot\vec{ds}$. Let $\vec{dr}$ be the line segment joining the two points (x,t) and $(x+dx,t+dt)$ (see Fig. 2.2(a)). Let $\vec{n}=(n_x,n_t)$ be a unit normal to $\vec{dr}$. Then

$$n_x = \pm\frac{dt}{\sqrt{(dx)^2+(dt)^2}} \quad , \quad n_t = \mp\frac{dx}{\sqrt{(dx)^2+(dt)^2}} \tag{2.7}$$

where $(dx)^2+(dt)^2 \neq 0$ is assumed. The upper and lower signs in Eq. (2.7) correspond to the two senses of $\vec{n}$. Let $\vec{ds}$ be the surface element with the end points (x,t) and $(x+dx,t+dt)$. Then

$$\vec{ds} \stackrel{def}{=} \sqrt{(dx)^2+(dt)^2}\;\vec{n} = \pm(dt\,,\,-dx) \tag{2.8}$$

Eqs. (2.4) and (2.8) imply that

$$\vec{\underline{h}}\cdot\vec{ds} = \pm\left[\left(a\underline{u}-\mu\frac{\partial \underline{u}}{\partial x}\right)dt-\underline{u}\,dx\right] = \pm\vec{\underline{e}}\cdot\vec{dr} \tag{2.9}$$

where

$$\vec{\underline{e}} \stackrel{def}{=} \left(-\underline{u}\,,\,a\underline{u}-\mu\frac{\partial \underline{u}}{\partial x}\right) \quad , \quad \vec{dr} \stackrel{def}{=} (dx\,,\,dt) \tag{2.10}$$

It may be shown that the upper (lower) signs in Eqs. (2.7) - (2.9) should be chosen if the 90° rotation from $\vec{n}$ to $\vec{dr}$ is in the counterclockwise (clockwise) direction.

Let Γ be a simple closed curve in E_2. Let (x,t) and $(x+dx,t+dt)$ be two points on Γ. Let $\vec{n}$ be the outward normal to Γ at the point (x,t) (see Fig. 2.2(b)). Then the upper (lower) signs in Eqs. (2.7) - (2.9) should be chosen if $\vec{dr}$ points in the counterclockwise (clockwise) direction of Γ. Let $\Delta\Gamma$ be a segment of Γ. Then Eq. (2.9) implies that

$$\int_{\Delta\Gamma}\vec{\underline{h}}\cdot\vec{ds} = \int_{\Delta\Gamma}^{c.c.}\vec{\underline{e}}\cdot\vec{dr} \tag{2.11}$$

where the notation *c.c.* indicates that the line integration should be carried out in the counterclockwise direction.

Let ◊PQRS be the parallelogram depicted in Fig. 2.1(b). Let

$$J(\overline{PQ}) \stackrel{def}{=} \text{the flux of } \vec{h} \text{ leaving } \Diamond PQRS \text{ through the line segment } \overline{PQ}. \tag{2.12}$$

Similarly, we define $J(\overline{QR})$, $J(\overline{RS})$ and $J(\overline{SP})$. Let Γ be the boundary of ◊PQRS. Then Eqs. (2.6), (2.10) and (2.11) may be used to obtain

$$J(\overline{PQ}) = \frac{\Delta x}{2}\left[(1+\tau)\gamma_j^n + (1-\tau^2-\delta)\frac{\Delta x}{4}\alpha_j^n\right] \tag{2.13}$$

$$J(\overline{QR}) = \frac{\Delta x}{2}\left[(1-\tau)\gamma_j^n - (1-\tau^2-\delta)\frac{\Delta x}{4}\alpha_j^n\right] \tag{2.14}$$

$$J(\overline{RS}) = \frac{\Delta x}{2}\left[-(1+\tau)\gamma_j^n + (1-\tau^2+\delta)\frac{\Delta x}{4}\alpha_j^n\right] \tag{2.15}$$

$$J(\overline{SP}) = \frac{\Delta x}{2}\left[-(1-\tau)\gamma_j^n - (1-\tau^2+\delta)\frac{\Delta x}{4}\alpha_j^n\right] \tag{2.16}$$

where

$$\tau \stackrel{def}{=} \frac{(a-b)\Delta t}{\Delta x} \tag{2.17}$$

and

$$\delta \stackrel{def}{=} \frac{4\mu\Delta t}{(\Delta x)^2} \geq 0 \tag{2.18}$$

Two comments may be made about Eqs. (2.13) - (2.16):

a. These equations are consistent with the local conservation relation

$$J(\overline{PQ}) + J(\overline{QR}) + J(\overline{RS}) + J(\overline{SP}) = 0 \tag{2.19}$$

and

b. The influence of parameters a and b on the fluxes leaving ◊PQRS through its four edges is expressed through a single parameter τ. As a result, the use of the moving mesh depicted in Fig. 2.1(a) does not increase the complexity of the expressions on the right sides of Eqs. (2.13) - (2.16). The meaning of τ is a subject to be discussed in Section 3.

To proceed, let

$$f_1^{(o)}(j,n) \stackrel{def}{=} \frac{2}{\Delta x}J(\overline{PQ}) \quad , \quad f_2^{(o)}(j,n) \stackrel{def}{=} \frac{2}{\Delta x}J(\overline{QR}) \tag{2.20}$$

$$f_1^{(I)}(j,n) \stackrel{def}{=} -\frac{2}{\Delta x} J(\overline{RS}) \quad , \quad f_2^{(I)}(j,n) \stackrel{def}{=} -\frac{2}{\Delta x} J(\overline{SP}) \tag{2.21}$$

In other words, $f_1^{(O)}(j,n)$ and $f_2^{(O)}(j,n)$, respectively, are the *normalized* fluxes *leaving* CE(j,n) through its "future right" and "future left" edges. Similarly, $f_1^{(I)}(j,n)$ and $f_2^{(I)}(j,n)$, respectively, are the *normalized* fluxes *entering* CE(j,n) through its "past left" and "past right" edges. For simplicity, a normalized flux will be referred to simply as a flux. Thus, the first two normalized fluxes may be referred to as the outgoing fluxes while the last two normalized fluxes as the incoming fluxes. These two pairs of fluxes form two column matrices, i.e.,

$$\vec{f}^{(O)}(j,n) \stackrel{def}{=} \begin{bmatrix} f_1^{(O)}(j,n) \\ f_2^{(O)}(j,n) \end{bmatrix} \quad , \quad \vec{f}^{(I)}(j,n) \stackrel{def}{=} \begin{bmatrix} f_1^{(I)}(j,n) \\ f_2^{(I)}(j,n) \end{bmatrix} \tag{2.22}$$

Also we define

$$q_1(j,n) \stackrel{def}{=} \gamma_j^n \quad , \quad q_2(j,n) \stackrel{def}{=} \frac{\Delta x}{4} \alpha_j^n \tag{2.23}$$

and

$$\vec{q}(j,n) \stackrel{def}{=} \begin{bmatrix} q_1(j,n) \\ q_2(j,n) \end{bmatrix} \tag{2.24}$$

With the aid of Eqs. (2.20) - (2.24), Eqs. (2.13) - (2.16) can be rewritten as

$$\vec{f}^{(O)}(j,n) = \Lambda^{(O)} \vec{q}(j,n) \tag{2.25}$$

and

$$\vec{f}^{(I)}(j,n) = \Lambda^{(I)} \vec{q}(j,n) \tag{2.26}$$

where $\Lambda^{(O)}$ and $\Lambda^{(I)}$ are the matrices defined by

$$\Lambda^{(O)} \stackrel{def}{=} \begin{bmatrix} 1+\tau & 1-\tau^2-\delta \\ 1-\tau & -(1-\tau^2-\delta) \end{bmatrix} \tag{2.27}$$

and

$$\Lambda^{(I)} \stackrel{def}{=} \begin{bmatrix} 1+\tau & -(1-\tau^2+\delta) \\ 1-\tau & 1-\tau^2+\delta \end{bmatrix} \tag{2.28}$$

Through out this paper, it will be assumed that

$$1 - \tau^2 + \delta \neq 0 \tag{2.29}$$

Thus $[\Lambda^{(I)}]^{-1}$, the inverse of $\Lambda^{(I)}$, exists. We have

$$[\Lambda^{(I)}]^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -\dfrac{1-\tau}{1-\tau^2+\delta} & \dfrac{1+\tau}{1-\tau^2+\delta} \end{bmatrix} \tag{2.30}$$

From Eq. (2.26), one obtains

$$\vec{q}(j,n) = [\Lambda^{(I)}]^{-1}\vec{f}^{(I)}(j,n) \tag{2.31}$$

Substituting Eq. (2.31) into Eq. (2.25), one has

$$\vec{f}^{(O)}(j,n) = \Omega\vec{f}^{(I)}(j,n) \tag{2.32}$$

where Ω is the matrix defined by

$$\Omega \stackrel{def}{=} \Lambda^{(O)}[\Lambda^{(I)}]^{-1} \tag{2.33}$$

Let ω_{lm}, $l, m = 1, 2$, be the elements of the matrix Ω. Then Eqs. (2.27), (2.30) and (2.33) imply that

$$\omega_{11} = \frac{\tau(1-\tau^2)+\delta}{1-\tau^2+\delta} \quad , \quad \omega_{12} = \frac{(1+\tau)(1-\tau^2)}{1-\tau^2+\delta} \tag{2.34}$$

$$\omega_{21} = \frac{(1-\tau)(1-\tau^2)}{1-\tau^2+\delta} \quad , \quad \omega_{22} = \frac{-\tau(1-\tau^2)+\delta}{1-\tau^2+\delta} \tag{2.35}$$

A result of Eqs. (2.34) and (2.35) is

$$\sum_{l=1}^{2} \omega_{lm} = 1 \quad , \qquad m = 1, 2 \tag{2.36}$$

Since Eq. (2.32) is equivalent to

$$f_l^{(O)}(j,n) = \sum_{m=1}^{2} \omega_{lm} f_m^{(I)}(j,n) \quad , \qquad l = 1, 2 \tag{2.37}$$

Eq. (2.36) may be used to prove that

$$f_1^{(O)}(j,n) + f_2^{(O)}(j,n) = f_1^{(I)}(j,n) + f_2^{(I)}(j,n) \tag{2.38}$$

i.e., the sum of the outgoing fluxes is equal to the sum of the incoming fluxes. From Eqs. (2.20) and (2.21), it is easy to see that Eq. (2.38) is equivalent to Eq. (2.19).

With the above preliminaries, the current marching procedure may now be defined by using the interface flux conservation relation. Explicitly, this relation requires that (see Fig. 2.3)

$$f_1^{(I)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) = f_1^{(O)}(j,n) \quad , \quad f_2^{(I)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) = f_2^{(O)}(j+1,n) \tag{2.39}$$

and

$$f_1^{(I)}(j,n+1) = f_1^{(O)}(j-\tfrac{1}{2},n+\tfrac{1}{2}) \quad , \quad f_2^{(I)}(j,n+1) = f_2^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) \qquad (2.40)$$

where $j = 0, \pm 1, \pm 2, \cdots$; $n = 0, 1, 2, \cdots$. Because of the above relations, a single arrow is drawn across an interface (see Fig. 2.3) to represent both the flux entering and the flux leaving this interface. Let $n \geq 0$ be a fixed integer. Let the outgoing fluxes $f_l^{(O)}(j,n)$ of the conservation elements at the time level n be given. According to Eq. (2.39), the incoming fluxes $f_1^{(I)}(j+\tfrac{1}{2},n+\tfrac{1}{2})$ and $f_2^{(I)}(j+\tfrac{1}{2},n+\tfrac{1}{2})$ of CE$(j+\tfrac{1}{2},n+\tfrac{1}{2})$, respectively, are equal to the outgoing flux $f_1^{(O)}(j,n)$ of CE(j,n) and the outgoing flux $f_2^{(O)}(j+1,n)$ of CE$(j+1,n)$. Thus all the incoming fluxes of the conservation elements at time level $n+\tfrac{1}{2}$ are known. Since Eq. (2.37) remains valid if the indices j and n, respectively, are replaced by $j+\tfrac{1}{2}$ and $n+\tfrac{1}{2}$, one has

$$f_l^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) = \sum_{m=1}^{2} \omega_{lm} f_m^{(I)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) \quad , \qquad l = 1, 2 \qquad (2.41)$$

The outgoing fluxes of the conservation elements at time level $n+\tfrac{1}{2}$ can be evaluated in terms of the known incoming fluxes by using Eq. (2.41). Similarly, with the aid of Eq. (2.37) and (2.40), the incoming and outgoing fluxes of the conservation elements at time level $n+1$ can also be evaluated. This procedure can continue for time levels $n+3/2$, $n+2$, $\cdots$. The following comments are made to provide more details about this procedure:

a. The outgoing fluxes $f_l^{(O)}(j, 0)$ at time level $n = 0$ may be evaluated by using Eq. (2.25) if the coefficients $q_l(j, 0)$ at time level $n = 0$ are given.

b. The coefficients ω_{lm} are considered as given constants in the marching procedure. Let j and n be a pair of fixed integers. Then the outgoing fluxes $f_l^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2})$, $l = 1, 2$, may be evaluated in terms of the incoming fluxes $f_m^{(I)}(j+\tfrac{1}{2},n+\tfrac{1}{2})$, $m = 1, 2$, by using Eq. (2.41). This evaluation requires four multiplications and two additions. However, the operation count can be reduced if the following alternative procedure is adopted. Let $f_1^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2})$ be evaluated by using Eq. (2.41). This requires two multiplications and one addition. $f_2^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2})$ is then evaluated by using the conservation relation

$$f_2^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) = f_1^{(I)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) + f_2^{(I)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) - f_1^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) \qquad (2.42)$$

which can be obtained from Eq. (2.38) by replacing j and n, respectively, with $j+\tfrac{1}{2}$ and $n+\tfrac{1}{2}$. The last evaluation requires only one addition and one subtraction. Thus evaluation of $f_l^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2})$, $l = 1, 2$, requires only two multiplications, two additions, and one subtraction. Thus the operation count of the current scheme is five for each j per half time step.

c. The principal variables involved in the marching procedure described above are incoming and outgoing fluxes. However, for any pair of given integers or half integers j and n, $\vec{q}(j,n)$ can always be evaluated in terms of $\vec{f}^{(I)}(j,n)$ by using Eq. (2.31).

More relations connecting to the marching procedure may be derived. To proceed, note that Eqs. (2.39) and (2.40), respectively, can be written as

$$\vec{f}^{(I)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) = \mathrm{I}_+ \vec{f}^{(O)}(j,n) + \mathrm{I}_- \vec{f}^{(O)}(j+1,n) \tag{2.43}$$

and

$$\vec{f}^{(I)}(j,n+1) = \mathrm{I}_+ \vec{f}^{(O)}(j-\tfrac{1}{2},n+\tfrac{1}{2}) + \mathrm{I}_- \vec{f}^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) \tag{2.44}$$

where

$$\mathrm{I}_+ \stackrel{def}{=} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad , \text{ and } \quad \mathrm{I}_- \stackrel{def}{=} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.45}$$

are projection matrices [p.116, 5]. For any pair of numbers c_1 and c_2, we have

$$\mathrm{I}_+ \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ 0 \end{bmatrix} \quad , \quad \mathrm{I}_- \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ c_2 \end{bmatrix} \tag{2.46}$$

Substituting Eq. (2.43) into the equation obtained from Eq. (2.31) by replacing j and n, respectively, with $j+\frac{1}{2}$ and $n+\frac{1}{2}$, one obtains

$$\vec{q}(j+\tfrac{1}{2},n+\tfrac{1}{2}) = [\Lambda^{(I)}]^{-1} \left[\mathrm{I}_+ \vec{f}^{(O)}(j,n) + \mathrm{I}_- \vec{f}^{(O)}(j+1,n) \right] \tag{2.47}$$

With the aid of Eq. (2.25), Eq. (2.47) implies that

$$\vec{q}(j+\tfrac{1}{2},n+\tfrac{1}{2}) = \mathrm{Q}_+ \vec{q}(j,n) + \mathrm{Q}_- \vec{q}(j+1,n) \tag{2.48}$$

where

$$\mathrm{Q}_+ \stackrel{def}{=} [\Lambda^{(I)}]^{-1} \mathrm{I}_+ \Lambda^{(O)} \quad , \quad \mathrm{Q}_- \stackrel{def}{=} [\Lambda^{(I)}]^{-1} \mathrm{I}_- \Lambda^{(O)} \tag{2.49}$$

Multiplying Eq. (2.47) from the left by $\Lambda^{(O)}$ and using Eqs. (2.25) and (2.33), one has

$$\vec{f}^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) = \Omega_+ \vec{f}^{(O)}(j,n) + \Omega_- \vec{f}^{(O)}(j+1,n) \tag{2.50}$$

where

$$\Omega_+ \stackrel{def}{=} \Omega \mathrm{I}_+ = \begin{bmatrix} \omega_{11} & 0 \\ \omega_{21} & 0 \end{bmatrix} \quad , \quad \Omega_- \stackrel{def}{=} \Omega \mathrm{I}_- = \begin{bmatrix} 0 & \omega_{12} \\ 0 & \omega_{22} \end{bmatrix} \tag{2.51}$$

Similarly, with the aid of Eq. (2.44), one can obtain

$$\vec{q}(j,n+1) = \mathrm{Q}_+ \vec{q}(j-\tfrac{1}{2},n+\tfrac{1}{2}) + \mathrm{Q}_- \vec{q}(j+\tfrac{1}{2},n+\tfrac{1}{2}) \tag{2.52}$$

and

$$\vec{f}^{(O)}(j,n+1) = \Omega_+ \vec{f}^{(O)}(j-\tfrac{1}{2},n+\tfrac{1}{2}) + \Omega_- \vec{f}^{(O)}(j+\tfrac{1}{2},n+\tfrac{1}{2}) \tag{2.53}$$

As a result of Eqs. (2.48) and (2.52), one has

$$\vec{q}(j,n+1) = [\,Q_+\,]^2 \vec{q}(j-1,n) + [\,Q_+Q_- + Q_-Q_+\,]\vec{q}(j,n) + [\,Q_-\,]^2 \vec{q}(j+1,n) \tag{2.54}$$

Similarly, Eqs. (2.50) and (2.53) may be used to obtain

$$\vec{f}^{(O)}(j,n+1) = [\,\Omega_+\,]^2 \vec{f}^{(O)}(j-1,n) + [\,\Omega_+\Omega_- + \Omega_-\Omega_+\,]\vec{f}^{(O)}(j,n) + [\,\Omega_-\,]^2 \vec{f}^{(O)}(j+1,n) \tag{2.55}$$

In both Eqs. (2.54) and (2.55), a column matrix at time level $n+1$ is expressed directly in terms of three column matrices at time level n. Note that, with the aid of Eqs. (2.23), (2.27), (2.30), (2.45) and (2.49), Eq. (2.54) can be explicitly expressed as

$$\begin{aligned}
\gamma_j^{n+1} &= \frac{1}{2}\left[\tau + \frac{\delta(1-\tau)}{1-\tau^2+\delta}\right]\left[(1+\tau)\gamma_{j-1}^n + \frac{(1-\tau^2-\delta)}{4}\Delta x\,\alpha_{j-1}^n\right] \\
&+ \frac{1-\tau^2}{1-\tau^2+\delta}\left[(1-\tau^2)\gamma_j^n - \frac{\tau(1-\tau^2-\delta)}{4}\Delta x\,\alpha_j^n\right] \\
&- \frac{1}{2}\left[\tau - \frac{\delta(1+\tau)}{1-\tau^2+\delta}\right]\left[(1-\tau)\gamma_{j+1}^n - \frac{(1-\tau^2-\delta)}{4}\Delta x\,\alpha_{j+1}^n\right]
\end{aligned} \tag{2.56}$$

and

$$\begin{aligned}
\Delta x\,\alpha_j^{n+1} &= -\frac{1}{2}\left[\tau + \frac{\delta(1-\tau)}{1-\tau^2+\delta}\right]\left[4\frac{1-\tau^2}{1-\tau^2+\delta}\gamma_{j-1}^n + (1-\tau)\frac{1-\tau^2-\delta}{1-\tau^2+\delta}\Delta x\,\alpha_{j-1}^n\right] \\
&+ \left[\frac{1-\tau^2}{1-\tau^2+\delta}\right]^2\left[4\tau\gamma_j^n + (1-\tau^2-\delta)\Delta x\,\alpha_j^n\right] \\
&- \frac{1}{2}\left[\tau - \frac{\delta(1+\tau)}{1-\tau^2+\delta}\right]\left[4\frac{1-\tau^2}{1-\tau^2+\delta}\gamma_{j+1}^n - (1+\tau)\frac{1-\tau^2-\delta}{1-\tau^2+\delta}\Delta x\,\alpha_{j+1}^n\right]
\end{aligned} \tag{2.57}$$

At this juncture, it is noted that the marching procedure described earlier is constructed by using Eqs. (2.50) and (2.53). One may construct alternative procedures by using Eqs. (2.48) and (2.52), or Eq. (2.54), or Eq. (2.55). However, these alternatives are less efficient than the original procedure. This is because *(i)* the matrices Ω_+ and Ω_-, respectively, have only two surviving elements, and *(ii)* the original procedure can take advantage of the conservation relation Eq. (2.38). Since the current numerical method is developed on the principle of flux conservation, it is natural that the most efficient marching procedure is the one that uses incoming and outgoing fluxes as the marching variables.

To further clarify the differences between the current scheme and other explicit schemes, this section is concluded with the following remarks:

a. The marching step in the Lax-Wendroff scheme in which $u_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ is updated, i.e., Eq. (1.7), is derived with the assumption that u is a constant along a straight line with $dx/dt = a$. This assumption generally breaks down if u satisfies Eq. (2.2) instead of Eq. (1.3). If the diffusion term in Eq. (2.2) is comparable to the convection term, the error caused by this assumption may be substantial. Nevertheless, the marching step Eq. (1.7) or one of its variants is used in many generalized Lax-Wendroff schemes that are used to solve Eq. (2.2). This marching step generally is followed by another in which u_j^{n+1} is obtained by using the conservation relation (see Fig. 1.2(a))

$$u_j^{n+1}\,\Delta x - u_j^n\,\Delta x + \left[a\,u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - \mu\left(\frac{\partial u}{\partial x}\right)_{j+\frac{1}{2}}^{n+\frac{1}{2}} \right] \Delta t - \left[a\,u_{j-\frac{1}{2}}^{n+\frac{1}{2}} - \mu\left(\frac{\partial u}{\partial x}\right)_{j-\frac{1}{2}}^{n+\frac{1}{2}} \right] \Delta t = 0 \qquad (2.58)$$

where $\left(\frac{\partial u}{\partial x}\right)_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ and $\left(\frac{\partial u}{\partial x}\right)_{j-\frac{1}{2}}^{n+\frac{1}{2}}$, respectively, are the finite-difference approximations of $\partial u/\partial x$ at the mesh points $(j+\frac{1}{2}, n+\frac{1}{2})$ and $(j-\frac{1}{2}, n+\frac{1}{2})$. Generally, these approximations may be expressed in terms of the mesh values of u at the time level $n+\frac{1}{2}$.

b. The assumption that u is a constant along a straight line with $dx/dt = a$ may be avoided if one lets $u_{j+\frac{1}{2}}^{n+\frac{1}{2}}$, $j = 0, \pm 1, \pm 2, \cdots$, be lagged behind by one half time step, i.e., $u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = u_{j+\frac{1}{2}}^{n}$, $j = 0, \pm 1, \pm 2, \cdots$. Here $u_{j+\frac{1}{2}}^{n}$ may be obtained by interpolating the given values of u_j^n, $j = 0, \pm 1, \pm 2, \cdots$. Obviously, an explicit conservative scheme may be formed by combining this new assumption with Eq. (2.58).

c. The errors caused by the assumptions mentioned in (a) and (b) generally are considered as the penalty one pays for using an *explicit conservative time-accurate* scheme. One may avoid this penalty by using either an implicit scheme or an explicit nonconservative time-accurate scheme (e.g., the MacCormack scheme). The current scheme is an exception to the above common wisdom. *It is explicit, conservative and time-accurate, yet constructed without relying on the assumptions mentioned in* (a) *and* (b).

d. In Section 1, the pure-convection discrete conservation relation Eq. (1.9) was cast into an integral form (i.e., Eq. (1.14)) similar to the conservation law Eq. (1.1). For $\mu \neq 0$, one may be tempted to repeat the same feat by again assuming Eq. (1.12) but replacing Eq. (1.13) with

$$\vec{\underline{h}}(x,t) \stackrel{def}{=} \left(a\underline{u}(x,t) - \mu\frac{\partial \underline{u}(x,t)}{\partial x}, \underline{u}(x,t)\right) \tag{2.59}$$

However, because $\partial \underline{u}(x,t)/\partial x = 0$, Eq. (1.14) again is equivalent to Eq. (1.9). It will not be equivalent to a convection-diffusion conservation relation in the form of Eq. (2.58).

A desire to cast the discrete conservation relation into an integral form is a strong motive behind the current development. As a matter of fact, the integral form Eq. (1.18) is one of the basic building blocks of the current marching scheme. It is our belief that a conservative scheme that can be cast into an integral form not only is easier to interpret but also provides a more realistic simulation of the conservation laws.

e. In the current scheme, $u(x,t)$ is approximated by $\underline{u}(x,t)$ which is defined in Eq. (2.6). For $(x,t) \in CE''(j,n)$, $\underline{u}(x,t)$ is determined by two independent parameters γ_j^n and α_j^n which, respectively, represent u and $\partial u/\partial x$ at the point (x_j^n, t^n). The extra parameter α_j^n accorded to the current scheme allows a more precise specification of the discrete initial conditions. It also provides more leeway for $\underline{u}(x,t)$ to simulate a rapidly varying function $u(x,t)$, as often occurs across a shock or within a boundary layer.

f. According to Eqs. (2.32) and (2.33), the determination of $\vec{f}^{(O)}(j,n)$ in terms of $\vec{f}^{(I)}(j,n)$ requires the inversion of the matrix $\Lambda^{(I)}$. As a result, the current scheme is *locally implicit*. As will be shown in Section 4, the appearance of the factor $(1-\tau^2+\delta)$ in the denominators of two elements in $[\Lambda^{(I)}]^{-1}$ (see Eq. (2.30)) has a positive effect on stability.

g. Let

$$1 - \tau^2 - \delta \neq 0 \tag{2.60}$$

Then $[\Lambda^{(O)}]^{-1}$ and Ω^{-1} exist. Let

$$\tilde{\Omega}_+ \stackrel{def}{=} L_+ \Omega^{-1} \quad , \quad \tilde{\Omega}_- \stackrel{def}{=} L_- \Omega^{-1} \tag{2.61}$$

By using Eq. (2.32) and the interface flux conservation conditions, we obtain the time-reversal counterpart of Eq. (2.55), i.e.,

$$\begin{aligned}\vec{f}^{(O)}(j,n) = {} & [\tilde{\Omega}_-]^2 \vec{f}^{(O)}(j-1,n+1) \\ & + [\tilde{\Omega}_+\tilde{\Omega}_- + \tilde{\Omega}_-\tilde{\Omega}_+]\vec{f}^{(O)}(j,n+1) + [\tilde{\Omega}_+]^2 \vec{f}^{(O)}(j+1,n+1)\end{aligned} \tag{2.62}$$

Eq. (2.62) states that the discrete variables associated with a conservation element at time level n can be determined by those of its three closest conservation elements at time level $n+1$. Contrarily, in a typical classical scheme, e.g., the Lax-Wendroff scheme Eq. (1.11), a discrete variable at time level n may depend on all the discrete variables at time level $n+1$. Note there may be several solutions of $\vec{f}^{(O)}(j,n)$ for a given set of $\vec{f}^{(O)}(j-1,n+1)$,

$\vec{f}^{(O)}(j,n+1)$, $\vec{f}^{(O)}(j+1,n+1)$, if the current scheme is generalized to solve a nonlinear PDE, e.g., the Burgers' equation [p.154, 3].

3. THE DYNAMIC SPACE-TIME MESH

The main purpose of this section is to explore the concept of *a dynamic space-time mesh* and the need for *a unified treatment of physical variables and mesh parameters.* Specifically, we will demonstrate that stability and accuracy of a numerical calculation may be improved if the space-time mesh is allowed to evolve with the physical variables such that *the local convective motion of physical variables relative to the moving mesh is kept to a minimum.* To simplify the discussions, again we consider only Eq. (1.3) or Eq. (2.2). Also the coefficients a and μ, and the mesh parameters b, Δt, and Δx are assumed to be frozen at their local values.

The parameter τ defined in Section 2 plays a central role in the following discussions. As a result, its role as the Courant number for a moving mesh will be established immediately.

In Fig. 2.1(a), Q and S, respectively, denote the mesh points $(j,n+\frac{1}{2})$ and $(j,n-\frac{1}{2})$. The point T is on time level $n-\frac{1}{2}$ with $\overline{TQ}$ being in the direction of convection. Hereafter, by definition, a line segment in space-time is said to be in the direction of convection if $dx/dt = a$ along this line segment. Note that the direction of convection is identical to the characteristic direction of Eq. (2.2) only if $\mu = 0$. Points T and S are on the same time level and separated by a spatial distance $(a-b)\Delta t$. The parameter τ is the ratio between this distance and Δx. In the case where $b = 0$, i.e., the moving mesh is reduced to a stationary mesh, the spatial distance between T and S is reduced to $a\Delta t$. As a result, τ is reduced to the ordinary Courant number ν. For this reason, the parameter τ may be considered as the Courant number for a moving mesh.

To further explore the significance of τ, again we consider Eq. (1.3) and the Lax-Wendroff scheme which solves it. If the moving mesh depicted in Fig. 2.1(a) is used, then this scheme may be expressed as

$$u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{2}\left[(1+\tau)\, u_j^n + (1-\tau)\, u_{j+1}^n \right] \tag{3.1}$$

and

$$u_j^{n+1} = u_j^n - \tau\left[u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - u_{j-\frac{1}{2}}^{n+\frac{1}{2}} \right] \tag{3.2}$$

As in the derivation of Eqs. (1.7) and (1.10), Eq. (3.1) is obtained through the use of backward characteristic projection and linear interpolation while Eq. (3.2) represents a flux conservation relation over the parallelogram PUVR shown in Fig. 2.1(a). When $b = 0$, $\tau = \nu$ and Eqs. (3.1) and (3.2), respectively, are reduced to Eqs. (1.7) and (1.10). For this reason, the original Lax-Wendroff scheme may be viewed as a special case of the scheme defined by Eqs. (3.1) and (3.2). As will be shown later, a moving mesh relative to a coordinate system may become a stationary mesh relative to another coordinate system. As a result, a scheme defined by Eqs. (3.1) and (3.2) with any value of b may be reinterpreted as the original Lax-Wendroff scheme if it is viewed

from another coordinate system. This will provide an alternative (and more satisfying) proof for Eqs. (3.1) and (3.2).

Note that ν is a function of Δx, Δt, and a while τ is a function of Δx, Δt, a, and b. The extra independent variable b of the function τ corresponds to the extra degree of freedom introduced as a result of allowing the mesh to be "moving" relative to the coordinate system. It will be shown immediately that the time-step size limitation associated with the original Lax-Wendroff scheme may be removed by taking advantage of this added freedom.

According to the von Neumann analysis, the amplification factor of the scheme defined by Eqs. (3.1) and (3.2) is

$$A^{L.W.}(\theta) = 1 - \tau^2 (1 - \cos\theta) - i\tau \sin\theta \tag{3.3}$$

where θ is the phase angle variation in Δx of a plane-wave component. Eq. (3.3) implies that the stability condition is $|\tau| \leq 1$, i.e.,

$$\Delta t \leq \frac{\Delta x}{|a-b|} \qquad (a \neq b) \tag{3.4}$$

Let Δx be held constant. Then Eq. (3.4) implies that the stability bound for Δt becomes greater as $|a-b|$ becomes smaller. Since a and b, respectively, are the convection velocity of the physical variable u and the velocity of the moving mesh, $a-b$ is the velocity at which u is convected *relative* to the moving mesh. In this paper, $a-b$ and $|a-b|$, respectively, may simply be referred to as the relative convection velocity and the relative convection speed. As a result, one may say that the time step size limitation associated with the Lax-Wendroff scheme is due to the existence of a nonzero relative convection speed.

A large relative convection speed and thus a severe time-step size limitation, may result from an indiscriminate use of a stationary mesh. For the current case in which a is a constant, this limitation may be eliminated completely by using a moving mesh with $b = a$. *Even if a is a function of u, x, and t, the above discussion suggests that the time-step size limitation may be reduced sharply if the mesh is designed such that the local relative convection speed is kept to a minimum.*

At this juncture, we introduce another interpretation for the parameter τ, i.e., it is the product of the relative convection velocity $(a-b)$ and the mesh aspect ratio $(\Delta t/\Delta x)$. If only the stationary mesh (i.e., $b = 0$) is allowed, then τ can be made smaller only by making $\Delta t/\Delta x$ smaller, a move very costly in computational effort. On the other hand, if a moving mesh is allowed, τ may be made smaller by reducing $|a-b|$.

Next we study the dependency of accuracy on the relative convection speed $|a-b|$. Note that $A^{L.W.}(\theta) = 1$ when $\tau = 0$. Thus the numerical dissipation and dispersion vanish when $\tau = 0$

[pp. 93-94, 3]. Moreover, Eq. (3.2) implies that the numerical solution does not vary along a j mesh line when $\tau = 0$. Since a j mesh line is also a characteristic line of Eq. (1.3) when $\tau = 0$, the numerical solution coincides with the analytical solution at the mesh points. Since $\tau \to 0$ as $(a-b) \to 0$ when $\Delta t/\Delta x$ is held constant, the above observations suggest that the accuracy of the scheme defined by Eqs. (3.1) and (3.2) may also be improved by reducing $|a-b|$.

Since $A^{L.W.}(\theta) = e^{\mp i\theta}$ when $\tau = \pm 1$, the dissipation and dispersion also vanish when $\tau = \pm 1$. Again Eqs. (3.1) and (3.2) imply that the numerical solution is exact when $\tau = \pm 1$ (Note: The diagonal $\overline{RU}$ ($\overline{PV}$) of the parallelogram PUVR depicted in Fig. 2.1(a) is in the characteristic direction of Eq. (1.3) if $\tau = 1$ ($\tau = -1$)). Thus the numerical solution of Eqs. (3.1) and (3.2) will be highly accurate if one uses a mesh with $|\tau| < 1$ and $|\tau|$ is very close to 1 everywhere. However, since $|\tau| = 1$ is on the verge of instability, this strategy of obtaining accurate solutions may not be practical when the coefficient a is not a constant.

In order to further explore the significance of τ and $(a-b)$, in the following, we will study the transformation properties of several equations and parameters under a Galilean transformation. This study will also provide a systematic way to obtain the form of a classical finite-difference scheme over a moving mesh.

To proceed, we consider the Galilean transformation:

$$x' = x - b^* t \quad \text{and} \quad t' = t \tag{3.5}$$

where b^* is any real constant. Physically, (x',t') represents a coordinate system moving with the velocity b^* relative to the coordinate system (x,t). Assuming that the mesh is fixed in space-time, Eqs. (2.3) and (3.5) imply that the coordinates x' and t' for the mesh point (j,n) are given by

$$x' = x'^n_j \stackrel{def}{=} j\Delta x + n\, b' \Delta t \quad \text{and} \quad t' = t'^n \stackrel{def}{=} n\Delta t \tag{3.6}$$

where

$$b' \stackrel{def}{=} b - b^* \tag{3.7}$$

is the velocity of the moving mesh relative to the new coordinate system. In this paper, a parameter defined with respect to the new coordinate system is denoted by a prime. Immediately, we have

$$\Delta x' \stackrel{def}{=} x'^n_{j+1} - x'^n_j = \Delta x \quad , \quad \Delta t' \stackrel{def}{=} t'^{n+1} - t'^n = \Delta t \tag{3.8}$$

Also, Eq. (2.2) may be rewritten as

$$\frac{\partial u'}{\partial t'} + a' \frac{\partial u'}{\partial x'} - \mu' \frac{\partial^2 u'}{\partial x'^2} = 0 \tag{3.9}$$

where

$$a' \stackrel{def}{=} a - b^* \quad , \quad \mu' \stackrel{def}{=} \mu \tag{3.10}$$

and u' is a function of x' and t' such that

$$u'(x',t') = u(x,t) \tag{3.11}$$

An immediate result of Eqs. (3.7) and (3.10) is

$$a' - b' = a - b \tag{3.12}$$

Thus the relative convection velocity $(a-b)$ is invariant under the Galilean transformation Eq. (3.5). Also Eqs. (3.8), (3.10) and (3.12) imply that

$$\tau' \stackrel{def}{=} \frac{(a'-b')\Delta t'}{\Delta x'} = \frac{(a-b)\Delta t}{\Delta x} = \tau \quad , \quad \delta' \stackrel{def}{=} \frac{4\mu'\Delta t'}{(\Delta x')^2} = \frac{4\mu\Delta t}{(\Delta x)^2} = \delta \tag{3.13}$$

Moreover, it may be shown that, for any $(x,t) \in CE''(j,n)$

$$\underline{u}(x,t) = \underline{u}'(x',t') \stackrel{def}{=} \alpha'^n_j (x' - x'^n_j) + \beta'^n_j (t' - t'^n) + \gamma'^n_j \tag{3.14}$$

where

$$\alpha'^n_j \stackrel{def}{=} \alpha^n_j \quad , \quad \beta'^n_j \stackrel{def}{=} \beta^n_j + b^* \alpha^n_j \quad , \quad \gamma'^n_j \stackrel{def}{=} \gamma^n_j \tag{3.15}$$

With the aid of Eqs. (3.10) and (3.15), one concludes that $a'\alpha'^n_j + \beta'^n_j = 0$ if and only if $a\,\alpha^n_j + \beta^n_j = 0$.

From Eqs. (3.8), (3.13) and (3.15), one concludes that all the parameters and variables that appear in Eqs. (2.56) and (2.57) are invariant under the Galilean transformation Eq. (3.5). This property will be used to simplify the discussion given in Section 6.

Let $b^* = b$. Then $b' = 0$, and thus Eqs. (3.6), (3.8) and (3.13) imply that

$$x'^n_j = j\Delta x' \quad , \quad t'^n = n\Delta t' \qquad (b^* = b) \tag{3.16}$$

and

$$\tau = \tau' = \frac{a'\Delta t'}{\Delta x'} \qquad (b^* = b) \tag{3.17}$$

From Eqs. (3.16) and (3.17), one concludes that *(i)* the mesh is stationary relative to the primed coordinate system, and *(ii)* τ simply becomes an ordinary Courant number when it is expressed in terms of the primed parameters. Moreover, as a result of *(i)*, a classical finite-difference scheme

may now be expressed relative to this new coordinate system in its traditional form. As an example, the L/D-F scheme [p.161, 3] for solving Eq. (3.9) may be expressed as

$$\frac{u'^{n+1}_j - u'^{n-1}_j}{2\Delta t'} + a' \frac{u'^{n}_{j+1} - u'^{n}_{j-1}}{2\Delta x'} - \mu' \frac{u'^{n}_{j+1} + u'^{n}_{j-1} - u'^{n+1}_j - u'^{n-1}_j}{(\Delta x')^2} = 0$$

$$(b^* = b) \tag{3.18}$$

Since the mesh is fixed in space-time, Eq. (3.11) implies that $u'^{n}_{j} = u^{n}_{j}$ for any (j,n). With the aid of Eqs. (3.8) and (3.10), Eq. (3.18) may be rewritten as

$$\frac{u^{n+1}_j - u^{n-1}_j}{2\Delta t} + (a-b) \frac{u^{n}_{j+1} - u^{n}_{j-1}}{2\Delta x} - \mu \frac{u^{n}_{j+1} + u^{n}_{j-1} - u^{n+1}_j - u^{n-1}_j}{(\Delta x)^2} = 0 \tag{3.19}$$

This is the form of the L/D-F scheme when the mesh and coordinate system used are those depicted in Fig. 2.1(a). *As a result, when a stationary mesh ($b = 0$) is replaced by a moving mesh ($b \neq 0$) without changing the coordinate system, the only modification required in the form of the L/D-F scheme is to replace the coefficient a with $(a-b)$. This is also true for other classical schemes solving Eq. (1.3) or Eq. (2.2).* Note that the Courant number ν should be replaced by the parameter τ as the coefficient a is replaced by $(a-b)$. This is consistent with the fact that Eqs. (1.7) and (1.10), respectively, are converted into Eqs. (3.1) and (3.2) when ν is replaced by τ.

In conclusion, the previous discussions suggest that a reduction in the relative convection speed $|a-b|$ may improve stability and accuracy, and reduce dissipation and dispersion of numerical calculations. *Since the appearance of wiggles near a discontinuity is a result of numerical dispersion, the wiggles may also be reduced by reducing the relative convection speed.*

4. STABILITY, DISSIPATION AND DISPERSION

4.1 Preliminaries

In this section, the current numerical scheme will be studied using a discrete Fourier analysis. Specifically, we assume the initial periodic conditions: $\vec{q}(j,0) = \vec{q}(j+K,0)$, $j = 0, \pm 1, \pm 2, \cdots$, where K is an integer ≥ 3. With the aid of Eq. (2.54), by induction, it may be shown that $\vec{q}(j,n)$ is periodic at any time level n, i.e.,

$$\vec{q}(j,n) = \vec{q}(j+K,n) \qquad (j = 0, \pm 1, \pm 2, \cdots, n = 0, 1, 2, \cdots) \tag{4.1}$$

With the aid of Eqs. (2.54) and (4.1), $\vec{q}(j,n)$ can be expressed explicitly as a matrix function of j, n, K and the initial-value matrices $\vec{q}(l,0)$, $l = 1, 2, 3, \cdots, K-1$. The stability, dissipation and dispersion of $\vec{q}(j,n)$ are then studied by using this functional relation. According to Eqs. (2.25), (2.26) and (2.48), the other matrix variables, including $\vec{q}(j+\frac{1}{2},n+\frac{1}{2})$, $\vec{f}^{(I)}(j,n)$, $\vec{f}^{(O)}(j,n)$, $\vec{f}^{(I)}(j+\frac{1}{2},n+\frac{1}{2})$ and $\vec{f}^{(O)}(j+\frac{1}{2},n+\frac{1}{2})$, may be considered as functions of $\vec{q}(j,n)$. Their behaviors, therefore, may be inferred directly from those of $\vec{q}(j,n)$.

Since the current Fourier analysis also serves as the basis of an error analysis to be presented in Section 5, the following development will include materials that are needed there.

To proceed, let

$$\phi_j^{(k)} \stackrel{def}{=} \frac{1}{\sqrt{K}} \exp[2\pi i jk/K] \qquad i \equiv \sqrt{-1}$$
$$(j = 0, \pm 1, \pm 2, \cdots, k = 0, 1, 2, \cdots, K-1) \tag{4.2}$$

$\phi_j^{(k)}$ are periodic and orthonormal, i.e.,

$$\phi_j^{(k)} = \phi_{j+K}^{(k)} \qquad (j = 0, \pm 1, \pm 2, \cdots, k = 0, 1, 2, \cdots, K-1) \tag{4.3}$$

and

$$\sum_{l=0}^{K-1} \phi_l^{(k)} \overline{\phi_l^{(k')}} = \delta_{kk'} \qquad (k, k' = 0, 1, 2, \cdots, K-1) \tag{4.4}$$

where $\delta_{kk'}$ is the Kronecker delta symbol. As a result,

$$\vec{q}(j,n) = \sum_{k=0}^{K-1} \vec{q}(k,n)\, \phi_j^{(k)} \qquad (j = 0, \pm 1, \pm 2, \cdots, n = 0, 1, 2, \cdots) \tag{4.5}$$

where

$$\vec{q}(k,n) \stackrel{def}{=} \sum_{l=0}^{K-1} \vec{q}(l,n)\, \overline{\phi_l^{(k)}} \qquad (k = 0, 1, 2, \cdots, K-1, \; n = 0, 1, 2, \cdots) \tag{4.6}$$

Hereafter, unless specified otherwise, it is assumed that $k = 0, 1, 2, \cdots, K-1$; $j = 0, \pm 1, \pm 2, \cdots$; and $n = 0, 1, 2, \cdots$.

Furthermore, let

$$\theta_k \stackrel{def}{=} \begin{cases} 2\pi k/K & \text{if } K/2 \geq k \geq 0 \\ 2\pi(k-K)/K & \text{if } K-1 \geq k > K/2 \end{cases} \tag{4.7}$$

and

$$Q(\theta) \stackrel{def}{=} e^{-i\theta/2} Q_+ + e^{i\theta/2} Q_- \qquad (\pi \geq \theta > -\pi) \tag{4.8}$$

Note that θ_k, $k = 0, 1, 2, \cdots, K-1$, are deliberately defined such that

$$\pi \geq \theta_k > -\pi \tag{4.9}$$

Also, unless specified otherwise, hereafter we assume that $\pi \geq \theta > -\pi$. Substituting Eq. (4.5) into Eq. (2.54), and using Eqs. (4.3), (4.4), (4.7) and (4.8), one has

$$\vec{q}(k, n+1) = [Q(\theta_k)]^2 \vec{q}(k, n) \tag{4.10}$$

i.e., the amplification matrix for any k is the square of the matrix $Q(\theta_k)$. Combining Eqs. (4.2), (4.5) - (4.7) and (4.10), it may be shown that

$$\vec{q}(j, n) = \frac{1}{K} \sum_{k=0}^{K-1} [Q(\theta_k)]^{2n} \sum_{l=0}^{K-1} e^{i(j-l)\theta_k} \vec{q}(l, 0) \tag{4.11}$$

i.e., the matrices $\vec{q}(j, n)$ are determined uniquely if the initial-value matrices $\vec{q}(l, 0)$, $l = 0, 1, 2, \cdots, K-1$, are given.

With the aid of Eqs. (2.27), (2.30), (2.45), (2.49) and (4.8), it can be shown that

$$Q(\theta) = \begin{bmatrix} \cos(\theta/2) - i\tau\sin(\theta/2) & -i(1-\tau^2-\delta)\sin(\theta/2) \\ \dfrac{i(1-\tau^2)\sin(\theta/2)}{1-\tau^2+\delta} & -\dfrac{1-\tau^2-\delta}{1-\tau^2+\delta}[\cos(\theta/2) + i\tau\sin(\theta/2)] \end{bmatrix} \tag{4.12}$$

Let

$$\eta(\theta) \stackrel{def}{=} \delta\cos(\frac{\theta}{2}) - i\tau(1-\tau^2)\sin(\frac{\theta}{2}) \tag{4.13}$$

Then the eigenvalues of $Q(\theta)$ are

$$\sigma_\pm(\theta) \stackrel{def}{=} \frac{\eta(\theta) \pm \sqrt{[\eta(\theta)]^2 + (1-\tau^2)^2 - \delta^2}}{1-\tau^2+\delta} \tag{4.14}$$

In this paper, the principal square root is defined such that

$$\frac{\pi}{2} \geq \textit{the phase angle of its polar form} > -\frac{\pi}{2}$$

By applying the von Neumann analysis to Eq. (2.54), it may be shown that the amplification factors are the eigenvalues of $[Q(\theta)]^2$, i.e., $[\sigma_+(\theta)]^2$ and $[\sigma_-(\theta)]^2$.

To proceed further, note that, for each θ, the matrix $Q(\theta)$ is either nondefective or defective [p.353, 6]. If $Q(\theta)$ is nondefective, the Jordan form $\hat{Q}(\theta)$ of $Q(\theta)$ and its powers $[\hat{Q}(\theta)]^2$, $[\hat{Q}(\theta)]^3$, $\cdots$ may be chosen as [p.362, 6]

$$[\hat{Q}(\theta)]^m = \begin{bmatrix} [\sigma_+(\theta)]^m & 0 \\ 0 & [\sigma_-(\theta)]^m \end{bmatrix}, \qquad m = 1, 2, 3, \cdots \tag{4.15}$$

On the other hand, if $Q(\theta)$ is defective, we have [p.362, 6]

$$\left.\begin{aligned} &\sigma_+(\theta) = \sigma_-(\theta) \\ &\text{and} \\ &[\hat{Q}(\theta)]^m = \begin{bmatrix} [\sigma_+(\theta)]^m & m[\sigma_+(\theta)]^{m-1} \\ 0 & [\sigma_-(\theta)]^m \end{bmatrix}, \qquad m = 1, 2, 3, \cdots \end{aligned}\right\} \tag{4.16}$$

According to Jordan's theorem [p.362, 6], for each θ ($\pi \geq \theta > -\pi$), there exists a nonsingular matrix $G(\theta)$ such that

$$Q(\theta) = G(\theta)\,\hat{Q}(\theta)\,[G(\theta)]^{-1} \tag{4.17}$$

Note that matrix $G(\theta)$ is not unique. It can be shown that a matrix $\tilde{G}(\theta)$ can also convert $Q(\theta)$ to $\hat{Q}(\theta)$ if and only if $\tilde{G}(\theta) = G(\theta)\Psi(\theta)$ where $\Psi(\theta)$ is *(i)* an arbitrary 2×2 nonsingular diagonal matrix if $Q(\theta)$ has two distinct eigenvalues, or *(ii)* an arbitray 2×2 nonsingular matrix if $Q(\theta)$ is a multiple of the identity matrix, or *(iii)* a 2×2 nonsingular upper-triangular matrix with identical diagonal elements if $Q(\theta)$ is defective. The above comments are useful in a later discussion.

Substituting Eq. (4.17) into Eq. (4.11), one arrives at

$$\vec{q}(j,n) = \sum_{k=0}^{K-1} e^{ij\theta_k}\, G(\theta_k)\,[\hat{Q}(\theta_k)]^{2n}\,\vec{c}_k \tag{4.18}$$

where the column matrices $\vec{c}_k$ are defined by

$$\vec{c}_k \overset{def}{=} \frac{1}{K}\,[G(\theta_k)]^{-1} \sum_{l=0}^{K-1} e^{-il\theta_k}\,\vec{q}(l,0) \tag{4.19}$$

Note that there is a one-to-one relation between the set of column matrices $\vec{c}_k$, $k = 0, 1, 2, \cdots, K-1$, and the set of column matrices $\vec{q}(j,0)$, $j = 0, 1, 2, \cdots, K-1$. As a matter of fact, one has

$$\vec{q}(j,0) = \sum_{k=0}^{K-1} e^{ij\theta_k} G(\theta_k) \vec{c}_k \tag{4.20}$$

Let

$$G(\theta) = \begin{bmatrix} g_{11}(\theta) & g_{12}(\theta) \\ g_{21}(\theta) & g_{22}(\theta) \end{bmatrix} \tag{4.21}$$

and

$$\vec{g}_+(\theta) \stackrel{def}{=} \begin{bmatrix} g_{11}(\theta) \\ g_{21}(\theta) \end{bmatrix} \quad , \quad \vec{g}_-(\theta) \stackrel{def}{=} \begin{bmatrix} g_{12}(\theta) \\ g_{22}(\theta) \end{bmatrix} \tag{4.22}$$

With the aid of Eqs. (4.15), (4.16), (4.21) and (4.22), Eq. (4.17), which is equivalent to $Q(\theta)G(\theta) = G(\theta)\hat{Q}(\theta)$, implies that *(i)* $\vec{g}_+(\theta)$ is an eigenvector of $Q(\theta)$ with the eigenvalue $\sigma_+(\theta)$, *(ii)* $\vec{g}_-(\theta)$ is an eigenvector of $Q(\theta)$ with the eigenvalue $\sigma_-(\theta)$ if $Q(\theta)$ is nondefective, and *(iii)* $Q(\theta)\vec{g}_-(\theta) = \vec{g}_+(\theta) + \sigma_-(\theta)\vec{g}_-(\theta)$ if $Q(\theta)$ is defective.

Let $Q(\theta_k)$, $k = 0, 1, 2, \cdots, K-1$, be nondefective or defective with $\sigma_+(\theta_k) = \sigma_-(\theta_k) = 0$ (Note: $[Q(\theta_k)]^2 = 0$ in the latter case). Then Eq. (4.18) is reduced to

$$\vec{q}(j,n) = \sum_{k=0}^{K-1} e^{ij\theta_k} \{ [\sigma_+(\theta_k)]^{2n} c_{k+} \vec{g}_+(\theta_k) + [\sigma_-(\theta_k)]^{2n} c_{k-} \vec{g}_-(\theta_k) \} \tag{4.23}$$

where c_{k+} and c_{k-}, respectively, are the upper and lower elements of the column matrix $\vec{c}_k$. Several comments may be made relating to Eq. (4.23):

a. The influence of the initial-value matrices $\vec{q}(l,0)$ on $\vec{q}(j,n)$ is expressed through the coefficients c_{k+} and c_{k-}.

b. Let

$$\vec{q}_\pm(j,n,k) \stackrel{def}{=} e^{ij\theta_k} [\sigma_\pm(\theta_k)]^{2n} c_{k\pm} \vec{g}_\pm(\theta_k) \tag{4.24}$$

Then, for each k, $\vec{q}(j,n) = \vec{q}_+(j,n,k)$ or $\vec{q}(j,n) = \vec{q}_-(j,n,k)$ is a particular solution of Eqs. (2.54) and (4.1). The general solution given in Eq. (4.23) is the sum of these particular solutions.

c. With the aid of Eqs. (2.45), (4.19), (4.21) and (4.22), and the fact that c_{k+} and c_{k-}, respectively, are the upper and lower elements of $\vec{c}_k$, it can be shown that

$$c_{k\pm} \vec{g}_\pm(\theta_k) = G(\theta_k) I_\pm \vec{c}_k = \frac{1}{K} G(\theta_k) I_\pm [G(\theta_k)]^{-1} \sum_{l=0}^{K-1} e^{-il\theta_k} \vec{q}(l,0) \tag{4.25}$$

Let the eigenvalues of $Q(\theta_k)$ be distinct. Then, as noted earlier, matrix $G(\theta_k)$ which converts $Q(\theta_k)$ into $\hat{Q}(\theta_k)$ (in this case, $\hat{Q}(\theta_k)$ is a diagonal matrix) can be replaced by

$\tilde{Q}(\theta_k) = G(\theta_k)\Psi(\theta_k)$ where $\Psi(\theta_k)$ is an arbitrary 2×2 nonsingular diagonal matrix. Since I_+, I_-, $\Psi(\theta_k)$ and $[\Psi(\theta_k)]^{-1}$ are diagonal and thus commute among themselves,

$$\tilde{G}(\theta_k)\, I_\pm\, [\tilde{G}(\theta_k)]^{-1} = G(\theta_k)\,\Psi(\theta_k)\, I_\pm\, [\Psi(\theta_k)]^{-1}\, [G(\theta_k)]^{-1} = G(\theta_k)\, I_\pm\, [G(\theta_k)]^{-1} \quad (4.26)$$

Combining Eqs. (4.25) and (4.26), one concludes that the matrices $c_{k\pm}\vec{g}_\pm(\theta_k)$ are invariants under the transformation $G(\theta_k) \rightarrow \tilde{G}(\theta_k)$ if the eigenvalues of $Q(\theta_k)$ are distinct.

To interpret the particular solutions defined in Eq. (4.24), we introduce the functions $\beta_\pm(\theta)$ such that

$$[\sigma_\pm(\theta)]^2 = |\sigma_\pm(\theta)|^2\, e^{i\beta_\pm(\theta)} \qquad \text{and} \quad \pi \geq \beta_\pm(\theta) > -\pi \quad (4.27)$$

Note that $\beta_+(\theta)$ ($\beta_-(\theta)$) is uniquely defined by Eq. (4.27) if $\sigma_+(\theta) \neq 0$ ($\sigma_-(\theta) \neq 0$). Also we define

$$\underline{a}_\pm(\theta) \stackrel{def}{=} b - \frac{\beta_\pm(\theta)}{\theta}\frac{\Delta x}{\Delta t} \qquad (\theta \neq 0) \quad (4.28)$$

$$\underline{\mu}_\pm(\theta) \stackrel{def}{=} -\frac{\ln|\sigma_\pm(\theta)|^2}{(\frac{\theta}{\Delta x})^2 \Delta t} \qquad (\theta \neq 0) \quad (4.29)$$

and

$$p_\pm(x,t,\theta) \stackrel{def}{=} \begin{cases} e^{-\underline{\mu}_\pm(\theta)(\frac{\theta}{\Delta x})^2 t + i(\frac{\theta}{\Delta x})[x - \underline{a}_\pm(\theta)t]} & \text{if } \theta \neq 0 \\ |\sigma_\pm(0)|^{2\frac{t}{\Delta t}}\, e^{i\beta_\pm(0)\frac{t}{\Delta t}} & \text{if } \theta = 0 \end{cases} \quad (4.30)$$

By using Eqs. (2.3) and (4.27) - (4.30), it may be shown that

$$e^{ij\theta_k}\, [\sigma_\pm(\theta_k)]^{2n} = p_\pm(x_j^n, t^n, \theta_k) \quad (4.31)$$

According to Eqs. (4.24) and (4.31), $\vec{q}_\pm(j,n,k)$ is the product of $p_\pm(x_j^n, t^n, \theta_k)$ and $c_{k\pm}\vec{g}_\pm(\theta_k)$. Since the latter is independent of j and n, the behaviors of $\vec{q}_\pm(j,n,k)$ are governed by the former.

For any θ such that $\pi \geq \theta > -\pi$ and $\theta \neq 0$, $u = p_\pm(x,t,\theta)$ is a plane-wave solution of the convection-diffusion equation

$$\frac{\partial u}{\partial t} + \underline{a}_\pm(\theta)\frac{\partial u}{\partial x} - \underline{\mu}_\pm(\theta)\frac{\partial^2 u}{\partial x^2} = 0$$

Also, the wavelength of this solution is given by

$$\lambda(\theta) \stackrel{def}{=} \frac{2\pi\Delta x}{|\theta|} \quad (4.32)$$

As a result of the above observations, for each $k \neq 0$ (i.e., $\theta_k \neq 0$), the particular solution $\vec{q}(j,n) = \vec{q}_{\pm}(j,n,k)$ may be referred to as the plane-wave solution with the numerical covection speed $\underset{\sim}{a}_{\pm}(\theta_k)$, the numerical viscosity $\underset{\sim}{\mu}_{\pm}(\theta_k)$ and the wavelength $\lambda(\theta_k)$. Also since $\theta_0 = 0$ and $p_{\pm}(x,t,0)$ is independent of x, one may say that the particular solution $\vec{q}(j,n) = \vec{q}_{\pm}(j,n,0)$ has an infinitely long wavelength.

In this paper, the marching procedure defined by Eqs. (2.54) and (4.1) is said to be stable if and only if, for any integer $K \geq 3$ and any specification of the matrices $\vec{c}_k$, $k = 0, 1, 2, \cdots, K-1$ (i.e., any specification of the initial-value matrices $\vec{q}(l,0)$. See Eqs. (4.19) and (4.20)), the elements of the column matrices $\vec{q}(j,n)$, $j = 0, \pm 1, \pm 2, \cdots$, remain bounded as $n \rightarrow +\infty$ with the parameters τ and δ being held constant (i.e., Δt and Δx being held constant — if one assumes that a, b and μ are constants). The readers are reminded that the term "stability" referred to in Lax's equivalence theorem has a meaning different from what we define here (see Section 6).

Because $G(\theta_k)$, $k = 0, 1, 2, \cdots, K-1$, are nonsingular, Eq. (4.18) implies that the marching procedure is stable if and only if, for any integer $K \geq 3$, the elements of the matrices $[\hat{Q}(\theta_k)]^{2n}$, $k = 0, 1, 2, \cdots, K-1$, remain bounded as $n \rightarrow +\infty$ with the parameters τ and δ being held constant. According to Eqs. (4.15) and (4.16), this implies that stability occurs if and only if, for any $K \geq 3$ and any $k = 0, 1, 2, \cdots, K-1$,

$$\max \{ \, |\sigma_+(\theta_k)| \, , \, |\sigma_-(\theta_k)| \, \} \leq 1 \quad \text{if } Q(\theta_k) \text{ is nondefective} \tag{4.33}$$

and

$$|\sigma_+(\theta_k)| < 1 \quad \text{if } Q(\theta_k) \text{ is defective} \tag{4.34}$$

Our study of dissipation and dispersion will be limited to the case in which each matrix $Q(\theta_k)$ is either nondefective or defective with $\sigma_+(\theta_k) = \sigma_-(\theta_k) = 0$. From Eqs. (4.23), (4.24), and (4.27) - (4.32), one concludes that *(i)* for any $k \neq 0$, the dissipation of $\vec{q}_{\pm}(j,n,k)$ may be measured by $\underset{\sim}{\mu}_{\pm}(\theta_k)$, and *(ii)* the dispersion of the general solution $\vec{q}(j,n)$ may be determined by the distribution of $\underset{\sim}{a}_{\pm}(\theta_k)$, $k = 1, 2, \cdots, K-1$.

As a final note of this subsection, we will point out a remarkable similarity between the forms of $\sigma_{\pm}(\theta)$ and the amplification factors $A_{\pm}(\theta)$ (see Eq. (B.10)) of the L/D-F scheme. Let $1 - \tau^2 > 0$. Then both the numerator and denominator on the right side of Eq. (4.14) may be divided by $1 - \tau^2$. As a result, we have

$$\sigma_{\pm}(\theta) = \frac{\hat{\delta}\cos(\frac{\theta}{2}) - i\tau\sin(\frac{\theta}{2}) \pm \sqrt{[\hat{\delta}\cos(\frac{\theta}{2}) - i\tau\sin(\frac{\theta}{2})]^2 + 1 - \hat{\delta}^2}}{1 + \hat{\delta}} \tag{4.35}$$

where $\hat{\delta} \overset{def}{=} \delta/(1-\tau^2)$. A comparison between Eqs. (4.35) and (B.10) reveals that the expression

on the right side of Eq. (B.10) may be converted to that on the right side of Eq. (4.35) if $\delta/2$, τ and θ, respectively, are replaced by $\hat{\delta}$, τ and $\theta/2$. In making this comparison, the reader should keep in mind that the amplification factors for the current scheme are $[\sigma_+(\theta)]^2$ and $[\sigma_-(\theta)]^2$, rather than $\sigma_+(\theta)$ and $\sigma_-(\theta)$. Note that, if $1-\tau^2 < 0$, the sign "$\pm$" on the right side of Eq. (4.35) should be replaced by "$\mp$".

This completes the preliminaries. A discussion of two special cases, i.e., *(i)* $\delta = 0$ and *(ii)* $\tau = 0$, will precede the investigation into the general case in which both δ and τ may not vanish.

4.2 The Special Case With $\delta = 0$

Eqs. (2.29) and (4.14) coupled with the assumption $\delta = 0$ imply that *(i)* $\tau^2 \neq 1$, and *(ii)*

$$\sigma_\pm(\theta) = -i\,\tau\sin(\theta/2) \pm \frac{|1-\tau^2|}{1-\tau^2}\sqrt{1-\tau^2\sin^2(\theta/2)} \tag{4.36}$$

In the following, Eq. (4.36) will be used to study *(a)* stability and dissipation, and *(b)* dispersion.

(a) Stability and Dissipation:

We have

$$|\sigma_\pm(\theta)| = \begin{cases} 1 & \text{if } |\tau\sin(\theta/2)| \le 1 \\ \left| -\tau\sin(\theta/2) \pm \dfrac{|1-\tau^2|}{1-\tau^2}\sqrt{\tau^2\sin^2(\theta/2)-1}\,\right| & \text{if } |\tau\sin(\theta/2)| > 1 \end{cases} \tag{4.37}$$

In the case where $\tau^2 > 1$, there exist a K and a k ($K-1 \ge k \ge 0$) such that $|\tau\sin(\theta_k/2)| > 1$. Thus

$$\max\{\,|\sigma_+(\theta_k)|\,,\,|\sigma_-(\theta_k)|\,\} > |\tau\sin(\theta_k/2)| > 1 \tag{4.38}$$

Combining Eqs. (4.33), (4.34), and (4.38), one concludes that the current marching procedure is not stable if $\tau^2 > 1$ and $\delta = 0$.

Let $\tau^2 < 1$. The Eq. (4.36) implies that $\sigma_+(\theta) \neq \sigma_-(\theta)$ for any θ. As a result, the matrices $Q(\theta)$, $\pi \ge \theta > -\pi$, are nondefective. According to Eq. (4.37), we also have $|\sigma_\pm(\theta)| = 1$, $\pi \ge \theta > -\pi$. It follows from Eq. (4.33) that the marching procedure is stable if $\tau^2 < 1$ and $\delta = 0$. Moreover, since $|\sigma_\pm(\theta_k)| = 1$, $k = 0, 1, 2, \cdots, K-1$, the particular solutions defined by Eq. (4.24) will not dissipate as n increases. Thus the numerics reflects faithfully the physics of pure convection. This contrasts sharply with the Lax-Wendroff scheme Eq. (1.11) which is numerically diffusive even though it is also a conservative scheme.

(b) Dispersion:

Let $\tau^2 < 1$. Then Eq. (4.36) implies that

$$[\sigma_{\pm}(\theta)]^2 = e^{\mp 2i \mathrm{Sin}^{-1}[\tau \sin(\theta/2)]} \tag{4.39}$$

Since $\theta_0 = 0$ and $[\sigma_{\pm}(0)]^2 = 1$, the particular solutions defined in Eq. (4.24) are independent of j and n if $k = 0$. As a matter of fact, the term with $k = 0$ on the right side of Eq. (4.23) is reduced to a constant column matrix $c_{0+}\vec{g}_+(0) + c_{0-}\vec{g}_-(0)$. Thus this term is ignored in the following discussion of the dispersion of the general solution defined in Eq. (4.23).

Since the range of Sin^{-1} is $(-\pi/2, \pi/2]$, one concludes that

$$\frac{\pi}{2} > \mathrm{Sin}^{-1}[\tau \sin(\theta/2)] > -\frac{\pi}{2} \qquad \text{if } \tau^2 < 1 \tag{4.40}$$

As a result, a comparison between Eqs. (4.27) and (4.39) reveals that

$$\pi > \beta_{\pm}(\theta) = \mp 2\, \mathrm{Sin}^{-1}[\tau \sin(\theta/2)] > -\pi \tag{4.41}$$

Substituting Eq. (4.41) into Eq. (4.28) and using Eq. (2.17), one concludes that

$$\underset{\sim}{a}_{\pm}(\theta) = a - \left\{ \tau \mp \frac{\mathrm{Sin}^{-1}[\tau \sin(\theta/2)]}{\theta/2} \right\} \frac{\Delta x}{\Delta t} \qquad (\theta \neq 0) \tag{4.42}$$

Thus

$$\underset{\sim}{a}_{\pm}(\theta) = a \qquad \text{if } \tau = 0 \text{ and } \theta \neq 0 \tag{4.43}$$

i.e., when $\tau = 0$, all the particular solutions which appear on the right side of Eq. (4.23) except the one with $k = 0$ are "convected" with the same velocity a. In this case, dispersion is completely absent.

In the case where $1 > \tau^2 > 0$, Eq. (4.42) may be expressed as

$$\underset{\sim}{a}_{\pm}(\theta) = a - \tau [\, 1 \mp F(\tau,\theta)\,] \frac{\Delta x}{\Delta t} \qquad (1 > \tau^2 > 0;\ \theta \neq 0) \tag{4.44}$$

where

$$F(\tau,\theta) \overset{def}{=} \frac{\mathrm{Sin}^{-1}(\tau \sin(\theta/2))}{\tau(\theta/2)} \qquad (1 > \tau^2 > 0;\ \theta \neq 0) \tag{4.45}$$

It is an exercise in calculus to show that

$$1 > F(\tau,\theta) > 2/\pi \tag{4.46}$$

By using Eqs. (4.44) and (4.46), one obtains that

$$0 < [\, a - \underset{\sim}{a}_+(\theta)\,] / (\tau \frac{\Delta x}{\Delta t}) < 1 - \frac{2}{\pi} \qquad (1 > \tau^2 > 0;\ \theta \neq 0) \tag{4.47}$$

and

$$1 + \frac{2}{\pi} < [\, a - \underset{\sim}{a}_-(\theta)\,] / (\tau \frac{\Delta x}{\Delta t}) < 2 \qquad (1 > \tau^2 > 0;\ \theta \neq 0) \tag{4.48}$$

Eqs. (4.47) and (4.48) state that, on the real line, the distance between any $\underset{\sim}{a}_+(\theta)$ and the physical convection velocity a is less than $(1 - \frac{2}{\pi})|\tau|\Delta x/\Delta t$ while the distance between any $\underset{\sim}{a}_-(\theta)$ and a is greater than $(1 + \frac{2}{\pi})|\tau|\Delta x/\Delta t$ and less than $2|\tau|\Delta x/\Delta t$. Thus the dispersion, measured by the maximum spread between a and any $\underset{\sim}{a}_+(\theta)$ or $\underset{\sim}{a}_-(\theta)$, is less than $2|\tau|\Delta x/\Delta t$. As $|\tau|$ decreases, so does the dispersion. Recall that the same conclusion was also reached in Section 3 for the Lax-Wendroff scheme. Moreover, since the maximum of the spread between a and $\underset{\sim}{a}_+(\theta)$ is less than the minimum of the spread between a and any $\underset{\sim}{a}_-(\theta)$, a particular solution defined by taking the upper sign in Eq. (4.24) will be "convected" at a velocity closer to the physical convection velocity a than a particular solution defined by taking the lower sign in Eq. (4.24). For this reason $\sigma_+(\theta)$ and $\vec{g}_+(\theta)$, respectively, may be referred to as the principal eigenvalue and eigenvector of matrix $Q(\theta)$ while $\sigma_-(\theta)$ and $\vec{g}_-(\theta)$, respectively, the spurious eigenvalue and eigenvector of $Q(\theta)$. This designation may be extended to the case $\theta = 0$ even though $\underset{\sim}{a}_\pm(\theta)$ are undefined at $\theta = 0$. Similarly, a particular solution $\vec{q}(j,n) \stackrel{def}{=} \vec{q}_\pm(j,n,k)$ will be referred to as a principal (spurious) solution if the upper (lower) sign is chosen.

4.3 The Special Case With $\tau = 0$

For this special case, the physical variable u has no convective motion relative to the moving mesh. Also Eq. (4.14) is reduced to

$$\sigma_\pm(\theta) = \frac{\delta\cos(\theta/2) \pm \sqrt{1 - \delta^2\sin^2(\theta/2)}}{1 + \delta} \tag{4.49}$$

Eq. (4.49), coupled with *(i)* $\delta \geq 0$ and *(ii)* $\cos(\theta/2) \geq 0$ if $\pi \geq \theta > -\pi$, implies that

$$|\sigma_\pm(\theta)|^2 = \begin{cases} \left[\dfrac{\delta\cos(\theta/2) \pm \sqrt{1 - \delta^2\sin^2(\theta/2)}}{1 + \delta} \right]^2 \leq 1 & \text{if } |\delta\sin(\theta/2)| \leq 1 \\ \dfrac{\delta - 1}{\delta + 1} < 1 & \text{if } |\delta\sin(\theta/2)| \geq 1 \end{cases} \tag{4.50}$$

Moreover, with the aid of Eq. (4.27) and the additional definitions that *(i)* $\beta_+(\theta) \stackrel{def}{=} 0$ if $\sigma_+(\theta) = 0$ and *(ii)* $\beta_-(\theta) \stackrel{def}{=} 0$ if $\sigma_-(\theta) = 0$, one concludes that

$$\beta_{\pm}(\theta) = \begin{cases} 0 & \text{if } |\delta \sin(\theta/2)| \le 1 \\ \pm 2\,\mathrm{Sin}^{-1}\left[\sqrt{\dfrac{\delta^2\sin^2(\theta/2)-1}{\delta^2-1}}\right] & \text{if } |\delta \sin(\theta/2)| > 1 \end{cases} \tag{4.51}$$

To study the stability, note that $\sigma_+(\theta) = \sigma_-(\theta)$ is a necessary condition for the defectiveness of matrix $Q(\theta)$, and it occurs if and only if $|\delta \sin(\theta/2)| = 1$. As a result, it follows from Eqs. (4.33), (4.34), and (4.50) that *the current scheme is unconditionally stable if* $\tau = 0$.

The dissipation and dispersion of the particular solutions defined by Eq. (4.24) generally may be studied explicitly by using Eqs. (4.28), (4.29), (4.50), and (4.51). This study is greatly simplified if one considers only the case in which $1 \ge \delta \ge 0$. For this special case, $|\delta \sin(\theta/2)| \le 1$ for any θ. According to Eq. (4.51), $\beta_{\pm}(\theta) = 0$, $\pi \ge \theta > -\pi$, i.e., the dispersion is absent. Moreover, by studying the extrema of the first expression on the right side of Eq. (4.50), it may be shown that

$$1 \ge |\sigma_+(\theta)|^2 \ge \frac{1-\delta}{1+\delta} \ge |\sigma_-(\theta)|^2 \ge \left(\frac{1-\delta}{1+\delta}\right)^2 \qquad (1 \ge \delta \ge 0) \tag{4.52}$$

According to Eqs. (4.24) and (4.52), the rate of dissipation per time step of any spurious solution is greater than or equal to that of any principal solution if $1 \ge \delta \ge 0$ and $\tau = 0$.

4.4 The General Case

Assuming $\delta \ge 0$ and $1 - \tau^2 + \delta \ne 0$, it is shown in Appendix A that the current scheme is stable if and only if $\tau^2 \le 1$. This stability condition has the remarkable property that it is independent of μ except that $\mu \ge 0$ is assumed.

Assuming $\delta \ge 0$, it is shown in Appendix B that the stability region of the L/D-F scheme on the $\delta - \tau$ plane is the region defined by $\tau^2 \le 1$, minus the two points $(0,1)$ and $(0,-1)$. *This stability region is exactly identical to that of the current scheme.*

On the other hand, the stability conditions of all other classical schemes known to the authors are dependent on μ. As an example, the stability region of the MacCormack scheme (see Appendix D) is depicted in Fig. 4.1. Obviously, the stability region of the MacCormack scheme is smaller than that of the L/D-F and the current scheme. The significance of this difference was discussed in Section 1. It will be further studied in sections 5 and 7.

The dissipation and dispersion properties of the current scheme may also be studied for the general case in which both δ and τ may not vanish. This requires the use of Eqs. (4.27) - (4.29) and (A.10).

5. ERROR ANALYSIS

In this section, an error analysis technique is developed using the discrete Fourier analysis formulated in Section 4. Assuming smooth initial data, this technique enables us to predict, analyze and compare the numerical errors of the L/D-F, the MacCormack, and the current schemes for calculations involving hundreds or thousands of time steps. As will be shown, the results of this error analysis provide us with a theoretical basis for improving the accuracy of the current scheme. They will also be used to interpret the numerical results to be presented in Section 7.

As a preliminary, the error analysis will be preceded by a discussion on a notable feature of the current scheme, i.e., the requirement to specify two sets of initial data involving the values of γ_j^0 and α_j^0.

Among the classical schemes, the initiation of the L/D-F scheme also requires the input of two sets of initial data, i.e., u_j^0 and u_j^1. Since only u_j^0 are given, generally u_j^1 are evaluated in terms of u_j^0 by using a starting condition, e.g., Eq. (B.1). Since the starting scheme is constructed with the aid of an one-sided difference approximation of a time derivative, it is one order less accurate than the main scheme. As a result, the accuracy of the L/D-F scheme may not attain the level that one would expect if only the main scheme is considered.

In the current scheme, the initial data γ_j^0 and α_j^0, respectively, will be identified with u_j^0 and $(\partial u/\partial x)_j^0$. In the case where $u(x, 0)$ is smooth and known for all x on the initial line, both u_j^0 and $(\partial u/\partial x)_j^0$ may be evaluated and used as the initial data for the numerical calculation. Generally, the extra set of initial data $(\partial u/\partial x)_j^0$ will allow a more accurate approximation of $u(x, 0)$ and thus gives the current method an edge in obtaining more accurate numerical solutions.

In the current error analysis, the accuracy of the MacCormack, the L/D-F and the current methods will be studied and compared assuming that only the initial data u_j^0 are given. For the current method, this means that $(\partial u/\partial x)_j^0$ must be evaluated in terms of u_j^0. Since $\partial u/\partial x$ at any point on the initial line $t = 0$ is the result of the differentiation of u *along* the initial line, $(\partial u/\partial x)_j^0$ may be expressed in terms of u_j^0 without using a one-sided difference approximation. Thus, at least in principle, the accuracy of the current scheme may not be reduced as a result of the complication associated with the extra initial data $(\partial u/\partial x)_j^0$. This contention will be verified by the results of the following analysis.

To proceed, let the initial data u_j^0, $j = 0, 1, 2, \cdots, K-1$ be given. Any u_j^0, $j = 0, \pm 1, \pm 2, \cdots$, may then be determined by using the periodic condition $u_j^0 = u_{j+K}^0$, $j = 0, \pm 1, \pm 2, \cdots$. In view of Eq. (2.6), it is natural to assume that

$$\gamma_j^0 = u_j^0 \quad , \quad j = 0, \pm 1, \pm 2, \cdots \tag{5.1}$$

In order to determine $(\partial u/\partial x)_j^0$ in terms of u_j^0, and also provide the initial values for a corresponding analytical problem (i.e., $u(x, 0)$ for this analytical problem will be determined in terms of u_j^0), a smooth periodic function $I(x)$ will be formed by linearly combining K periodic exponential functions (see Eq. (5.5)) such that

$$(a) \qquad I(j\Delta x) = u_j^0 \quad , \quad j = 0, 1, 2, \cdots, K-1 \tag{5.2}$$

and

$$(b) \qquad I(x+K\Delta x) = I(x) \tag{5.3}$$

As a result of Eq. (5.2), and the fact that α_j^0 is the spatial derivative in CE''$(j, 0)$, we will assume that

$$\alpha_j^0 = \dot{I}(j\Delta x) \quad , \quad j = 0, \pm 1, \pm 2, \cdots \tag{5.4}$$

where $\dot{I}(x)$ is the derivative of $I(x)$ with respect to x.

Given γ_j^0 and α_j^0, the discrete solution to Eqs. (2.54) and (4.1) may be determined. Assuming $u(x, 0) = I(x)$, the analytical solution to Eq. (2.2) with the periodic condition $u(x+K\Delta x, t) = u(x, t)$ may also be determined. The accuracy of the discrete solution may then be assessed by comparing it with the analytical solution.

To construct $I(x)$, note that the exponential functions

$$I_l(x) \stackrel{def}{=} e^{i2\pi lx/(K\Delta x)} \quad , \quad l = 0, \pm 1, \pm 2, \cdots \tag{5.5}$$

form a basis for the function space of the functions that have period $K\Delta x$ and are of bounded variation over $[0, K\Delta x]$ [p.478, 2]. Let the integer $K^* \geq 1$ be defined by

$$K^* \stackrel{def}{=} \begin{cases} (K-1)/2 & \text{if } K \text{ is odd} \\ K/2 & \text{if } K \text{ is even} \end{cases} \tag{5.6}$$

Then it may be shown that the wavelengths of the K functions $I_l(x)$, $l = K^*-K+1, K^*-K+2, \cdots, K^*$, are longer than or equal to those of the other functions defined in Eq. (5.5). We assume that $I(x)$ is a linear combination of these K functions. With the aid of Eqs. (4.7) and (5.2), it may be shown that

$$I(x) = \sum_{k=0}^{K-1} b_k e^{i\frac{\theta_k}{\Delta x}x} \tag{5.7}$$

where

$$b_k \stackrel{def}{=} \frac{1}{K} \sum_{l=0}^{K-1} e^{-il\theta_k} u_l^0 \quad , \quad k = 0, 1, 2, \cdots, K-1 \tag{5.8}$$

At this point, it should be emphasized that the function $I(x)$ defined by Eqs. (5.7) and (5.8) generally is complex even if the given initial data u_j^0 are real. As a result, generally α_j^0 may be complex. This should not cause alarm since the discrete equation to be solved, i.e., Eq. (2.54), is a system of algebraic equations with real constant coefficients. For these equations, both real and imaginary parts of a complex solution are themselves solutions. As a result, in case that physics so dictates, only the real parts of the initial data and solution may be considered as physically relevant. Obviously, the above comments are also applicable to a differential equation with real constant coefficients like Eq. (2.2).

To obtain the solution to Eqs. (2.54) and (4.1), note that a result of Eqs. (2.23), (2.24), (4.2), (4.4), (4.7), (5.1), (5.4), (5.7), and (5.8) is

$$\frac{1}{K} \sum_{l=0}^{K-1} e^{-il\theta_k} \vec{q}(l, 0) = b_k \begin{bmatrix} 1 \\ \frac{i\theta_k}{4} \end{bmatrix} \tag{5.9}$$

Combining Eqs. (4.25) and (5.9), one obtains

$$c_{k\pm} \vec{g}_{\pm}(\theta_k) = b_k \vec{H}_{\pm}(\theta_k) \quad , \quad k = 0, 1, 2, \cdots, K-1 \tag{5.10}$$

where

$$\vec{H}_{\pm}(\theta) \stackrel{def}{=} G(\theta) \mathrm{I}_{\pm} [G(\theta)]^{-1} \begin{bmatrix} 1 \\ \frac{i\theta}{4} \end{bmatrix} \tag{5.11}$$

By definition, $(\mathrm{I}_+ + \mathrm{I}_-)$ is the 2×2 identity matrix. Thus Eq. (5.11) implies that

$$\vec{H}_+(\theta) + \vec{H}_-(\theta) = \vec{H}_a(\theta) \stackrel{def}{=} \begin{bmatrix} 1 \\ \frac{i\theta}{4} \end{bmatrix} \tag{5.12}$$

In this section, we assume that every $Q(\theta_k)$, $k = 0, 1, 2, \cdots, K-1$, has two distinct eigenvalues. As a result, Eqs. (4.23) and (4.26) are applicable. With the aid of Eq. (5.10), Eqs. (4.23) and (4.24) imply that

$$\vec{q}(j,n) = \sum_{k=0}^{K-1} \vec{q}(j,n,k) \tag{5.13}$$

where

$$\vec{q}(j,n,k) \stackrel{def}{=} \vec{q}_+(j,n,k) + \vec{q}_-(j,n,k)$$

$$= b_k e^{ij\theta_k} \{ [\sigma_+(\theta_k)]^{2n} \vec{H}_+(\theta_k) + [\sigma_-(\theta_k)]^{2n} \vec{H}_-(\theta_k) \} \tag{5.14}$$

According to Eq. (5.14), $\vec{q}(j,n,k)$ is composed of the principal and the spurious parts. At $n = 0$, the amplitudes of these two parts are $b_k \vec{H}_+(\theta_k)$ and $b_k \vec{H}_-(\theta_k)$, respectively.

Let $u(x, 0) = I(x)$. Then the analytical solution to Eq. (2.2) with $u(x+K\Delta x,t) = u(x,t)$ is

$$u(x,t) = u_a(x,t) \stackrel{def}{=} \sum_{k=0}^{K-1} b_k e^{-\mu(\frac{\theta_k}{\Delta x})^2 t + i\frac{\theta_k}{\Delta x}(x-at)} \tag{5.15}$$

Let

$$\vec{q}_a(j,n) \stackrel{def}{=} \begin{bmatrix} u_a(x_j^n, t^n) \\ \frac{\Delta x}{4} \left[\frac{\partial u_a(x,t)}{\partial x} \right]_{x=x_j^n,\, t=t^n} \end{bmatrix} \tag{5.16}$$

In view of Eqs. (1.17a), (1.17b), (2.23), and (2.24), $\vec{q}_a(j,n)$ may be considered as the analytical counterpart of $\vec{q}(j,n)$. Let

$$A_a(\theta) \stackrel{def}{=} e^{-(i\tau + \frac{\delta\theta}{4})\theta} \tag{5.17}$$

and

$$\vec{q}_a(j,n,k) \stackrel{def}{=} b_k e^{ij\theta_k} [A_a(\theta_k)]^n \vec{H}_a(\theta_k) \tag{5.18}$$

Then Eqs. (5.12), (5.15), (5.16), (2.3), and (2.18) may be used to show that

$$\vec{q}_a(j,n) = \sum_{k=0}^{K-1} \vec{q}_a(j,n,k) \tag{5.19}$$

Combining Eqs. (5.12), (5.14), and (5.18), one obtains that

$$\Delta\vec{q}(j,n,k) \stackrel{def}{=} \vec{q}(j,n,k) - \vec{q}_a(j,n,k) = b_k e^{ij\theta_k} [\vec{E}_+(n,\theta_k) + \vec{E}_-(n,\theta_k)] \tag{5.20}$$

where

$$\vec{E}_\pm(n,\theta) \stackrel{def}{=} \{ [\sigma_\pm(\theta)]^{2n} - [A_a(\theta)]^n \} \vec{H}_\pm(\theta) \tag{5.21}$$

Thus $\vec{q}(j,n,k) = \vec{q}_a(j,n,k)$ if both $[\sigma_+(\theta)]^2$ and $[\sigma_-(\theta)]^2$ are replaced by $A_a(\theta)$. For this reason, $A_a(\theta)$ may be referred to as the analytical amplification factor. Because $\vec{q}_a(j,n)$ is the analytical counterpart of $\vec{q}(j,n)$, the numerical error of $\vec{q}(j,n)$ may be measured by

$$\Delta\vec{q}(j,n) \stackrel{def}{=} \vec{q}(j,n) - \vec{q}_a(j,n) = \sum_{k=0}^{K-1} \Delta\vec{q}(j,n,k) \tag{5.22}$$

The last equality sign follows from Eqs. (5.13), (5.19) and (5.20).

In the following, $\Delta\vec{q}(j,n,k)$ will be studied assuming

$$1 > \tau^2 \quad \text{and} \quad \delta \geq 0 \tag{5.23}$$

To proceed, we define

$$\omega(\theta) \stackrel{def}{=} \cos(\theta/2) - i\,\tau \sin(\theta/2) - \sigma_-(\theta) \quad , \qquad \pi \geq \theta > -\pi \tag{5.24}$$

In Appendix C, it is shown that *(i)* $\omega(\theta) \neq 0$ if $\pi > |\theta|$, and *(ii)* $\omega(\pi) \neq 0$ if either $1-\tau^2-\delta \neq 0$ or $0 > \tau > -1$. In the following discussion, we assume that $\omega(\theta) \neq 0$. Let

$$\xi_1(\theta) \stackrel{def}{=} \frac{(1-\tau^2)\sin(\theta/2)}{(1-\tau^2+\delta)\,\omega(\theta)} \quad , \quad \xi_2(\theta) \stackrel{def}{=} \frac{(1-\tau^2-\delta)\sin(\theta/2)}{\omega(\theta)} \tag{5.25}$$

Then it is shown in Appendix C that, for any θ such that $\pi \geq \theta > -\pi$ and $\sigma_+(\theta) \neq \sigma_-(\theta)$, one has

(*a*)
$$1 + \xi_1(\theta)\,\xi_2(\theta) \neq 0 \tag{5.26}$$

(*b*)
$$\vec{H}_+(\theta) = \frac{1 + \dfrac{\theta\,\xi_2(\theta)}{4}}{1 + \xi_1(\theta)\,\xi_2(\theta)} \begin{bmatrix} 1 \\ \\ i\,\xi_1(\theta) \end{bmatrix} \tag{5.27}$$

and

(*c*)
$$\vec{H}_-(\theta) = \frac{i\left[\dfrac{\theta}{4} - \xi_1(\theta)\right]}{1 + \xi_1(\theta)\,\xi_2(\theta)} \begin{bmatrix} i\,\xi_2(\theta) \\ \\ 1 \end{bmatrix} \tag{5.28}$$

Note that, for the special cases in which $\delta = 0$ or $\tau = 0$, $\xi_1(\theta)$ and $\xi_2(\theta)$ are given by

$$\xi_1(\theta) = \begin{cases} \dfrac{\sin(\theta/2)}{\cos(\theta/2) + \sqrt{1-\tau^2\sin^2(\theta/2)}} & \text{if } \delta = 0 \\ \\ \dfrac{\sin(\theta/2)}{\cos(\theta/2) + \sqrt{1-\delta^2\sin^2(\theta/2)}} & \text{if } \tau = 0 \end{cases} \tag{5.29}$$

and

$$\xi_2(\theta) = \begin{cases} \dfrac{(1-\tau^2)\sin(\theta/2)}{\cos(\theta/2)+\sqrt{1-\tau^2\sin^2(\theta/2)}} & \text{if } \delta = 0 \\ \dfrac{(1-\delta^2)\sin(\theta/2)}{\cos(\theta/2)+\sqrt{1-\delta^2\sin^2(\theta/2)}} & \text{if } \tau = 0 \end{cases} \tag{5.30}$$

In view of the different roles played by the parameters δ and τ, the structural similarity among the expressions on the right sides of Eqs. (5.29) and (5.30) is indeed remarkable.

According to Eq. (5.14) and the comment following Eq. (4.32), the wavelength of any particular solution $\vec{q}(j,n) = \vec{q}(j,n,k)$ is inversely proportional to $|\theta_k|$ if $k \neq 0$ and Δx is held constant. This is also true for its analytical counterpart. Thus a particular solution with a smaller $|\theta_k|$ is a slower-varying function of the index j. In the following, we will study $\Delta\vec{q}(j,n,k)$ assuming $|\theta_k|$ is small. Note that a general solution which is a slow-varying function of the index j generally is dominated by the particular solutions with small $|\theta_k|$, i.e., the coefficients b_k are very small except for those k's with very small $|\theta_k|$.

As a preliminary to the following use of the Taylor's expansion, note that, according to Eq. (4.14), the current assumption $\sigma_+(\theta) \neq \sigma_-(\theta)$ is valid if and only if

$$\zeta(\theta) \stackrel{def}{=} [\eta(\theta)]^2 + (1-\tau^2)^2 - \delta^2 \neq 0 \tag{5.31}$$

By assumption $1 > \tau^2$. Thus $\zeta(0) = (1-\tau^2)^2 > 0$ and $\omega(0) = 2(1-\tau^2)/(1-\tau^2+\delta) > 0$. It follows that that there is a neighborhood of $\theta = 0$ on the complex θ-plane in which both $\sqrt{\zeta(\theta)}$ and $\omega(\theta)$ are nonzero analytical functions of θ (Note: In obtaining the Taylor's expansions for the following study, the functions involved may be considered as the functions of a complex variable θ).

By using Eq. (5.28), it may be shown that

$$\vec{H}_-(\theta) = \begin{bmatrix} \dfrac{(1-\tau^2)^2-\delta^2}{32(1-\tau^2)}\left\{ \dfrac{i\tau\delta}{1-\tau^2}\theta^3 + \dfrac{1}{24}\left[(1+3\tau^2) + \dfrac{3\delta^2(1-9\tau^2)}{(1-\tau^2)^2}\right]\theta^4 + O(\theta^5)\right\} \\ \dfrac{\tau\delta\theta^2}{8(1-\tau^2)} - \dfrac{i}{192}\left[(1+3\tau^2) + \dfrac{3\delta^2(1-5\tau^2)}{(1-\tau^2)^2}\right]\theta^3 + O(\theta^4) \end{bmatrix} \tag{5.32}$$

Hereafter a quantity is denoted by $O(\theta^l)$ if there exists a constant $C > 0$ such that the absolute value of this quantity $\leq C|\theta|^l$ for all sufficiently small $|\theta|$. Eq. (5.32) is reduced to

$$\vec{H}_-(\theta) = \begin{bmatrix} \dfrac{(1-\tau^2)(1+3\tau^2)}{768}\theta^4 + O(\theta^5) \\ -\dfrac{i(1+3\tau^2)}{192}\theta^3 + O(\theta^4) \end{bmatrix} \quad \text{if } \delta = 0 \tag{5.33}$$

or

$$\vec{H}_-(\theta) = \begin{bmatrix} \dfrac{(1-\delta^2)(1+3\delta^2)}{768}\theta^4 + O(\theta^5) \\ -\dfrac{i(1+3\delta^2)}{192}\theta^3 + O(\theta^4) \end{bmatrix} \quad \text{if } \tau = 0 \tag{5.34}$$

Note that Taylor's expression of $\vec{H}_+(\theta)$ may be obtained from that of $\vec{H}_-(\theta)$ by using Eq. (5.12). An inspection of these expansions reveals that the upper and lower elements in $\vec{H}_-(\theta)$, respectively, are smaller than those in $\vec{H}_+(\theta)$ by *(i)* three orders and one order of θ if $\tau\delta \neq 0$ or *(ii)* four orders and two orders of θ if either $\delta = 0$ or $\tau = 0$. Thus, at n = 0, the spurious part of $\vec{q}(j,n,k)$ is much smaller than the principal part if $|\theta_k| \ll 1$.

In the case where $|\theta| \ll 1$, $\vec{H}_\pm(\theta)$ may be approximated by the leading terms in their Taylor's expansions. In the following, we will search for the approximations of

$$[\sigma_\pm(\theta)]^{2n} - [A_a(\theta)]^n$$

which are valid for small θ and *large* n. Note that the approximations which are valid for only small n are of little value since n generally is quite large in a typical numerical calculation.

To proceed, let

$$\varepsilon_+(\theta) \stackrel{def}{=} \frac{[\sigma_+(\theta)]^2}{A_a(\theta)} - 1 \tag{5.35}$$

By definition, $|\varepsilon_+(\theta)|$ is a measure of error when $A_a(\theta)$ is approximated by $[\sigma_+(\theta)]^2$. It may be shown that

$$\varepsilon_+(\theta) = \frac{i\tau\Delta_1(\tau,\delta)}{24}\theta^3 + \frac{\delta}{192}\left[\frac{1-5\tau^2}{1-\tau^2}\Delta_1(\tau,\delta) - 4\tau^2\right]\theta^4 + O(\theta^5) \tag{5.36}$$

where

$$\Delta_1(\tau,\delta) \stackrel{def}{=} 1 - \tau^2 - \frac{3\delta^2}{1-\tau^2} \tag{5.37}$$

is an often-encountered expression in the current paper.

Since

$$(i) \qquad [\sigma_+(\theta)]^{2n} - [A_a(\theta)]^n \equiv [A_a(\theta)]^n \{ [1+\varepsilon_+(\theta)]^n - 1 \} \tag{5.38}$$

and

$$(ii) \qquad [1+\varepsilon_+(\theta)]^n \doteq 1 + n\,\varepsilon_+(\theta) \quad \text{if } n|\varepsilon_+(\theta)| \ll 1 \tag{5.39}$$

we arrive at the important conclusion that

$$[\sigma_+(\theta)]^{2n} - [A_a(\theta)]^n \doteq n\,[A_a(\theta)]^n \varepsilon_+(\theta) \quad \text{if } n|\varepsilon_+(\theta)| \ll 1 \tag{5.40}$$

Here the sign "$\doteq$" is used to signal the fact that the ratio between the expressions on both sides of this sign is nearly equal to 1.

Note that the assumption $n|\varepsilon_+(\theta)| \ll 1$ may be justified if one is only interested in the case in which the numerical calculation is accurate up to the time step n, i.e., $\vec{q}(j,l,k) \doteq \vec{q}_a(j,l,k)$, $l = 1, 2, 3, \cdots, n$. According to Eqs. (5.20) and (5.21), generally this requires that $[\sigma_+(\theta_k)]^{2l} \doteq [A_a(\theta_k)]^l$, $l = 1, 2, \cdots, n$. It is shown in Appendix C that the last n equations are valid if and only if $n|\varepsilon_+(\theta_k)| \ll 1$.

Let

$$\varepsilon_-(\theta) \stackrel{def}{=} \frac{[\sigma_-(\theta)]^2}{\Delta_2(\tau,\delta)A_a(\theta)} - 1 \tag{5.41}$$

where

$$\Delta_2(\tau,\delta) \stackrel{def}{=} \left[\frac{1-\tau^2-\delta}{1-\tau^2+\delta}\right]^2 \tag{5.42}$$

By definition, $|\varepsilon_-(\theta)|$ is a measure of error when $\Delta_2(\tau,\delta)A_a(\theta)$ is approximated by $[\sigma_-(\theta)]^2$. It may be shown that

$$\varepsilon_-(\theta) = 2i\tau\theta + (\delta/2 - 2\tau^2)\theta^2 + O(\theta^3) \tag{5.43}$$

Eqs. (5.41) and (5.43) imply that $[\sigma_-(\theta)]^2 \to \Delta_2(\tau,\delta)A_a(\theta)$ as $|\theta| \to 0$.

Because *(i)* $\Delta_2(\tau,\delta) < 1$ if and only if $\delta(1-\tau^2) > 0$, and *(ii)* $\varepsilon_-(\theta) = O(\theta)$, Eq. (5.41) implies that

$$\left|\frac{[\sigma_-(\theta)]^2}{A_a(\theta)}\right| = |\Delta_2(\tau,\delta)[1+\varepsilon_-(\theta)]| < 1 \tag{5.44}$$

if $\delta > 0$ *and* $|\theta|$ *is sufficiently small* (note: $1-\tau^2 > 0$ is assumed in Eq. (5.23)). Assuming Eq. (5.44), one may conclude that $|[\sigma_-(\theta)]^{2n}/[A_a(\theta)]^n| \ll 1$ and thus

$$[\sigma_-(\theta)]^{2n} - [A_a(\theta)]^n \doteq -[A_a(\theta)]^n \tag{5.45}$$

if n is sufficiently large.

As an example, let $\tau = 0.5$ and $\delta = 0.1$. Then $\Delta_2(\tau,\delta) \doteq 0.585$. Let θ be sufficiently small such that $|\varepsilon_-(\theta)| < 0.1$. Then $|[\sigma_-(\theta)]^2 / A_a(\theta)| < 0.65$. Thus $|[\sigma_-(\theta)]^{2n} / [A_a(\theta)]^n| < 1.812 \times 10^{-4}$ if $n = 20$.

In the following discussions, the integers n and k are such that the approximations given in Eqs. (5.40) and (5.45) are valid if $\theta = \theta_k$. Let $b_k \neq 0$. Let $r_i(n,k)$, $i = 1, 2$, denote the ratio between the i-th element in the column matrix $\Delta\vec{q}(j,n,k)$ and that in the column matrix $\vec{q}_a(j,n,k)$ [Note: the meaning of $r_i(n,k)$ will be examined later]. Then Eqs. (5.18), (5.20) and (5.21) imply that

$$r_i(n,k) = r_{i+}(n,k) + r_{i-}(n,k) \quad , \quad i = 1, 2 \tag{5.46}$$

with

$$r_{i\pm}(n,k) \stackrel{def}{=} \frac{[\vec{E}_\pm(n,\theta_k)]_i}{[A_a(\theta_k)]^n [\vec{H}_a(\theta_k)]_i} \quad , \quad i = 1, 2 \tag{5.47}$$

Here $[\vec{E}_\pm(n,\theta_k)]_i$ and $[\vec{H}_a(\theta_k)]_i$, respectively, denote the i-th elements of the column matrices $\vec{E}_\pm(n,\theta_k)$ and $\vec{H}_a(\theta_k)$. Hereafter, $r_{i+}(n,k)$ and $r_{i-}(n,k)$, respectively, will be referred to as the principal and spurious parts of $r_i(n,k)$. By using Eqs. (5.12), (5.21), (5.32), (5.36), (5.40), and (5.45), one has

$$r_{1+}(n,k) \doteq n\left\{ \frac{i\tau\Delta_1(\tau,\delta)}{24}\theta_k^3 + \frac{\delta}{192}\left[\frac{1-5\tau^2}{1-\tau^2}\Delta_1(\tau,\delta) - 4\tau^2 \right]\theta_k^4 + O(\theta_k^5) \right\} \tag{5.48}$$

$$r_{1-}(n,k) \doteq -\frac{(1-\tau^2)^2 - \delta^2}{32(1-\tau^2)}\left\{ \frac{i\tau\delta}{1-\tau^2}\theta_k^3 + \frac{1}{24}\left[(1+3\tau^2) + \frac{3\delta^2(1-9\tau^2)}{(1-\tau^2)^2} \right]\theta_k^4 + O(\theta_k^5) \right\} \tag{5.49}$$

$$r_{2+}(n,k) \doteq n\left\{ \frac{i\tau\Delta_1(\tau,\delta)}{24}\theta_k^3 + \frac{\delta}{192}\left[\frac{1-9\tau^2}{1-\tau^2}\Delta_1(\tau,\delta) - 4\tau^2 \right]\theta_k^4 + O(\theta_k^5) \right\} \tag{5.50}$$

and

$$r_{2-}(n,k) \doteq \frac{i\tau\delta}{2(1-\tau^2)}\theta_k + \frac{1}{48}\left[(1+3\tau^2) + \frac{3\delta^2(1-5\tau^2)}{(1-\tau^2)^2} \right]\theta_k^2 + O(\theta_k^3) \tag{5.51}$$

Note that, in the current paper, any term denoted by $O(\theta^l)$ or $O(\theta_k^l)$ is independent of n.

Now we shall reexamine the meaning of $r_1(n,k)$ and $r_2(n,k)$. By using Eqs. (5.15), (5.17), (5.18), and (5.12), it may be shown that

$$u_a(x_j^n, t^n) = \sum_{k=0}^{K-1} u_a(j,n,k) \tag{5.52}$$

where

$$u_a(j,n,k) \stackrel{def}{=} b_k e^{ij\theta_k} [A_a(\theta_k)]^n = \text{ the upper element in } \vec{q}_a(j,n,k) \tag{5.53}$$

Let $\underline{u}(j,n,k)$ be the numerical counterpart of $u_a(j,n,k)$, i.e.,

$$\underline{u}(j,n,k) \stackrel{def}{=} \text{ the upper element in } \vec{q}(j,n,k) \tag{5.54}$$

Then it may be shown that

$$r_1(n,k) = \frac{\underline{u}(j,n,k) - u_a(j,n,k)}{u_a(j,n,k)} \qquad (b_k \neq 0) \tag{5.55}$$

Similarly, it may be shown that

$$u_{ax}(x_j^n, t^n) \stackrel{def}{=} \left[\frac{\partial u_a(x,t)}{\partial x} \right]_{x=x_j^n,\ t=t^n} = \sum_{k=0}^{K-1} u_{ax}(j,n,k) \tag{5.56}$$

where

$$u_{ax}(j,n,k) \stackrel{def}{=} \frac{i\theta_k}{\Delta x} b_k e^{ij\theta_k} [A_a(\theta_k)]^n$$

$$= \frac{4}{\Delta x} \times \left[\text{ the lower element in } \vec{q}_a(j,n,k) \right] \tag{5.57}$$

The numerical counterpart of $u_{ax}(j,n,k)$ is

$$\underline{u}_x(j,n,k) \stackrel{def}{=} \frac{4}{\Delta x} \times \left[\text{ the lower element in } \vec{q}(j,n,k) \right] \tag{5.58}$$

Obviously

$$r_2(n,k) = \frac{\underline{u}_x(j,n,k) - u_{ax}(j,n,k)}{u_{ax}(j,n,k)} \qquad (b_k \neq 0) \tag{5.59}$$

The accuracy of the current scheme will be compared with those of the L/D-F scheme and the MacCormack scheme. In Appendices B and D, we study the numerical solutions of Eq. (2.2) generated by these two schemes. In these studies, again we assume periodic conditions, i.e., Eq. (B.2) and use the mesh depicted in Fig. 2.1(a).

Let $\underline{u}_L(j,n,k)$ be the numerical counterpart of $u_a(j,n,k)$ obtained by using the L/D-F scheme (see Eqs. (B.31) and (B.33)). Let

$$r_L(n,k) \stackrel{def}{=} \frac{\underline{u}_L(j,n,k) - u_a(j,n,k)}{u_a(j,n,k)} \qquad (b_k \neq 0) \tag{5.60}$$

Then it is shown in Appendix B that

$$r_L(n,k) = r_{L+}(n,k) + r_{L-}(n,k) \tag{5.61}$$

with

$$r_{L+}(n,k) \doteq n\left\{ \frac{\tau^2 \delta}{4}\theta_k^2 + \frac{i\tau}{6}(1-\tau^2)(1-\frac{3}{4}\delta^2)\theta_k^3 \right.$$
$$\left. + \frac{\delta}{96}[9\tau^2\delta^2(1-\tau^2) + 3\tau^4\delta + 4\tau^2 - \frac{3}{2}\delta^2 + 2]\theta_k^4 + O(\theta_k^5) \right\} \tag{5.62}$$

and

$$r_{L-}(n,k) \doteq \frac{1}{4}(1-\frac{\delta^2}{4})\left\{ \tau^2\theta_k^2 - \frac{i\tau\delta}{2}(1-2\tau^2)\theta_k^3 \right.$$
$$\left. - \frac{1}{48}[4\tau^2(4-9\tau^2) + 3\delta^2(15\tau^4 - 12\tau^2 + 1)]\theta_k^4 + O(\theta_k^5) \right\} \tag{5.63}$$

$r_{L+}(n,k)$, and $r_{L-}(n,k)$, respectively, may be referred to as the principal and spurious parts of $r_L(n,k)$.

Let $\underline{u}_M(j,n,k)$ be the numerical counterpart of $u_a(j,n,k)$ obtained by using the MacCormack scheme. Let

$$r_M(n,k) \stackrel{def}{=} \frac{\underline{u}_M(j,n,k) - u_a(j,n,k)}{u_a(j,n,k)} \qquad (b_k \neq 0) \tag{5.64}$$

Then it is shown in Appendix D that

$$r_M(n,k) \doteq n\left\{ \frac{i\tau(1-\tau^2)}{6}\theta_k^3 + \frac{1}{48}[\delta(1-6\tau^2) - 6\tau^2(1-\tau^2)]\theta_k^4 + O(\theta_k^5) \right\} \tag{5.65}$$

The numerical errors of the current scheme, the L/D-F scheme, and the MacCormack scheme will be studied and compared using Eqs. (5.46), (5.48), (5.49), (5.61) - (5.63), and (5.65). Note that the parameter $r_2(n,k)$ and the associated equations, i.e., Eqs. (5.50) and (5.51), have no counterparts in the last two classical schemes.

According to Eqs. (5.48) and (5.49), the principal and the spurious parts of $r_1(n,k)$ are of the same order of θ_k. Because the principal part is linear in n while the spurious part is independent of n, generally, one expects that the principal part will become dominant as n increases. This conclusion also applies to $r_L(n,k)$. However, it may not apply to $r_2(n,k)$ because its principal part is $n\,O(\theta_k^3)$ while its spurious part is $O(\theta_k)$.

By comparing Eqs. (5.48), (5.49), (5.62), and (5.63), one concludes that *the principal and spurious parts of $r_1(n,k)$, respectively, are one order of θ_k smaller than those of $r_L(n,k)$.* As shown in Appendix B, the difference reflects the fact that *the current scheme is more accurate than the LID-F scheme in both the initial-value specification and the main marching procedure.*

According to Eqs. (5.48), (5.49), and (5.65), $r_M(n,k)$, and both the principal and the spurious parts of $r_1(n,k)$ are in the same order of θ_k. Moreover, one may observe that:

a. For any $\tau \neq 0$, the ratio between the leading term in the principal part of $r_1(n,k)$ and the leading term in $r_M(n,k)$ approaches 1/4 as $\delta \to 0$.

b. The leading term in the spurious part of $r_1(n,k)$ generally will be small for small δ.

The above observations coupled with the fact that the principal part would be dominant for a large n lead us to conclude *that the current scheme generally will be more accurate than the MacCormack scheme by a factor of 4 when δ is small and the initial condition is smooth.*

In the above discussion, the general constraints on τ and δ, respectively, are $1-\tau^2 > 0$ and $\delta > 0$. For the MacCormack scheme, these constraints must be further tightened by stability consideration (see Fig. 4.1). In the following, we will discuss the additional constraints required to annihilate the leading term in the principal part of each of the parameters $r_1(n,k)$, $r_2(n,k)$, $r_L(n,k)$, and $r_M(n,k)$ (Note: $r_M(n,k)$ has only the principal part). This annihilation obviously will lead to a sharp improvement in the accuracy of the schemes under consideration.

Let $\tau = 0$. Then all the leading terms in the principal and spurious parts of $r_1(n,k)$, $r_2(n,k)$, $r_L(n,k)$, and $r_M(n,k)$ vanish. As a matter of fact, it may be shown that

$$r_1(n,k) \doteq n\left[\frac{\delta(1-3\delta^2)}{192}\theta_k^4 + O(\theta_k^5)\right] - \frac{(1-\delta^2)(1+3\delta^2)}{768}\theta_k^4 + O(\theta_k^5) \tag{5.66}$$

$$r_2(n,k) \doteq n\left[\frac{\delta(1-3\delta^2)}{192}\theta_k^4 + O(\theta_k^5)\right] + \frac{(1+3\delta^2)}{48}\theta_k^2 + O(\theta_k^3) \tag{5.67}$$

$$r_L(n,k) \doteq n\left[\frac{\delta(4-3\delta^2)}{192}\theta_k^4 + O(\theta_k^5)\right] - \frac{\delta^2}{64}\left(1-\frac{\delta^2}{4}\right)\theta_k^4 + O(\theta_k^5) \tag{5.68}$$

and

$$r_M(n,k) \doteq n\left[\frac{\delta}{48}\theta_k^4 + O(\theta_k^5)\right] \tag{5.69}$$

Note that: *(i)* The second leading terms on the right sides of Eqs. (5.62) and (5.63) also vanish when $\tau = 0$, and *(ii)* when $\tau = 0$, the ratio of the leading terms in the remaining $r_{1+}(n,k)$, $r_{L+}(n,k)$, and $r_M(n,k)$ approaches 1:4:4 as $\delta \to 0$. Again we emphasize that, for a stationary mesh (i.e., $b = 0$), $\tau = 0$ occurs only when $a = 0$. However, for any $a \neq 0$, τ may be annihilated if one uses a moving mesh with $b = a$.

According to Eq. (5.65), the leading term in $r_M(n,k)$ also vanishes when $\tau^2 = 1$. However, since $\tau^2 = 1$ occurs either outside or on the stability boundary of the MacCormack scheme (see Fig. 4.1), the strategy of improving the accuracy of the MacCormack method by choosing the parameters Δt, Δx and b such that $\tau^2 \doteq 1$ may be impractical in reality. This is particularly true for the more general case in which the convection speed may be a function of the dependent variable u.

According to Eqs. (5.48) and (5.50), the leading terms in $r_{1+}(n,k)$ and $r_{2+}(n,k)$ also vanish when $\Delta_1(\tau,\delta) = 0$. Combining $\Delta_1(\tau,\delta) = 0$ and the assumption $1-\tau^2 > 0$, one obtains the optimal condition

$$1 - \tau^2 = \sqrt{3}\,\delta \tag{5.70}$$

Combining Eqs. (5.46), (5.48) - (5.51) and (5.70), it may be shown that

$$r_1(n,k) \doteq n\left[-\frac{\tau^2(1-\tau^2)}{48\sqrt{3}}\theta_k^4 + O(\theta_k^5)\right] - \frac{(1-\tau^2)}{48}\left[\frac{i\tau}{\sqrt{3}}\theta_k^3 + \frac{1}{12}(1-3\tau^2)\theta_k^4 + O(\theta_k^5)\right] \tag{5.71}$$

and

$$r_2(n,k) \doteq n\left[-\frac{\tau^2(1-\tau^2)}{48\sqrt{3}}\theta_k^4 + O(\theta_k^5)\right] + \frac{i\tau}{2\sqrt{3}}\theta_k + \frac{1}{24}(1-\tau^2)\theta_k^2 + O(\theta_k^3) \tag{5.72}$$

Eq. (5.70) represents a parabola on the $\delta-\tau$ plane. Since the segment of this parabola with $\delta > 0$ lies entirely within the stability boundary of the current method, no stability constraint may prevent us from improving the accuracy of the current method by choosing Δt, Δx, and b such that $1-\tau^2 = \sqrt{3}\,\delta$.

Because $\delta > 0$, Eq. (5.70) is equivalent to the parametric equations:

$$\tau = \tan(\psi/2) \qquad (\frac{\pi}{2} > \psi > -\frac{\pi}{2}) \tag{5.73}$$

and

$$\delta = \frac{1}{\sqrt{3}}[1 - \tan^2(\psi/2)] \qquad (\frac{\pi}{2} > \psi > -\frac{\pi}{2}) \tag{5.74}$$

Eqs. (5.73) and (5.74) imply that

$$\mathrm{Re} \stackrel{def}{=} \frac{\tau}{\delta} = \frac{\sqrt{3}}{2}\tan(\psi) \qquad (\frac{\pi}{2} > \psi > -\frac{\pi}{2}) \tag{5.75}$$

With the aid of Eqs. (2.17) and (2.18), one also has

$$\mathrm{Re} = \frac{(a-b)\Delta x}{4\mu} \tag{5.76}$$

As a result, Re may be referred to as the mesh Reynolds number.

In the case where $(a-b)$, μ and Δx are given, Re and ψ may be determined using Eqs. (5.76) and (5.75). Subsequently, the values of τ, δ, and Δt when $\Delta_1(\tau,\delta) = 0$ may be determined, respectively, by Eqs. (5.73) and (5.74), and the relation $\Delta t = (\Delta x)^2\,\delta/(4\mu)$. For later reference, these particular values of τ, δ and Δt, respectively, will be denoted by τ_0, δ_0 and $(\Delta t)_0$.

In the case where μ, Δt, and Δx are given, δ may be determined using Eq. (2.18). If $\delta > 1/\sqrt{3}$, Eq. (5.70) has no solution. If $1/\sqrt{3} \geq \delta > 0$, then Eq. (5.70) implies that $\Delta_1(\tau,\delta) = 0$ if

$$\tau = \tau_{\pm} \stackrel{def}{=} \pm\sqrt{1 - \sqrt{3}\,\delta} \tag{5.77}$$

Subsequently, the values of $(a-b)$ when $\Delta_1(\tau,\delta) = 0$ may be determined by using the relation $(a-b) = \tau\,\Delta x/\Delta t$.

Eqs. (5.49), (5.51), and (5.63) were obtained assuming Eqs. (5.45) and (B.43). The last two equations, in turn, assume $\delta > 0$. In the following, we consider the special case in which $\delta = 0$.

Eqs. (4.37) and (5.17) imply that

$$|\sigma_-(\theta)| = |A_a(\theta)| = 1 \qquad \text{if } \tau^2 < 1 \text{ and } \delta = 0 \tag{5.78}$$

As a result,

$$\left|\frac{[\sigma_-(\theta)]^{2n} - [A_a(\theta)]^n}{[A_a(\theta)]^n}\right| \leq 2 \qquad \text{if } \tau^2 < 1 \text{ and } \delta = 0 \tag{5.79}$$

With the aid of Eqs. (5.12), (5.33), (5.47), and (5.21), one concludes that

$$|r_{1-}(n,k)| \le \left| \frac{(1-\tau^2)(1+3\tau^2)}{384} \theta_k^4 + O(\theta_k^5) \right| \qquad (\delta = 0) \tag{5.80}$$

and

$$|r_{2-}(n,k)| \le \left| \frac{(1+3\tau^2)}{24} \theta_k^2 + O(\theta_k^3) \right| \qquad (\delta = 0) \tag{5.81}$$

Eqs. (5.48) and (5.50) are applicable even if $\delta = 0$. In that case, they, respectively, are reduced to

$$r_{1+}(n,k) \doteq n \left[\frac{i\tau(1-\tau^2)}{24} \theta_k^3 + O(\theta_k^5) \right] \qquad (\delta = 0) \tag{5.82}$$

and

$$r_{2+}(n,k) \doteq n \left[\frac{i\tau(1-\tau^2)}{24} \theta_k^3 + O(\theta_k^5) \right] \qquad (\delta = 0) \tag{5.83}$$

A comparison between Eqs. (5.80) and (5.82) reveals that the spurious part of $r_1(n,k)$ is negligible compared with its principal part if $\delta = 0$. Thus

$$r_1(n,k) \doteq r_{1+}(n,k) \doteq n \left[\frac{i\tau(1-\tau^2)}{24} \theta_k^3 + O(\theta_k^5) \right] \qquad (\delta = 0) \tag{5.84}$$

Since $r_{2-}(n,k) \le O(\theta_k^2)$ and $r_{2+}(n,k) \doteq n\, O(\theta_k^3)$, the relative weights of the principal and spurious parts of $r_2(n,k)$ are dependent on the relative magnitudes of θ_k and n if $\delta = 0$.

By using an argument similar to that leading to Eqs. (5.80) and (5.81), it is shown in Appendix B that

$$|r_{L-}(n,k)| \le \left| \frac{\tau^2}{2} \theta_k^2 - \frac{\tau^2(4-9\tau^2)}{24} \theta_k^4 + O(\theta_k^5) \right| \qquad (\delta = 0) \tag{5.85}$$

if $\tau^2 < 1$ and $\delta = 0$. Eq. (5.62) is applicable even if $\delta = 0$. In that case, it is reduced to

$$r_{L+}(n,k) \doteq n \left[\frac{i\tau(1-\tau^2)}{6} \theta_k^3 + O(\theta_k^5) \right] \qquad (\delta = 0) \tag{5.86}$$

Eqs. (5.85) and (5.86) imply that the relative weights of the principal and spurious parts of $r_L(n,k)$ are dependent on the relative magnitudes of θ_k and n if $\delta = 0$.

By comparing Eqs. (5.84), (5.86), and the reduced form of Eq. (5.65) for the special case $\delta = 0$, i.e.,

$$r_M(n,k) \doteq n\left[\frac{i\tau(1-\tau^2)}{6}\theta_k^3 - \frac{\tau^2(1-\tau^2)}{8}\theta_k^4 + O(\theta_k^5)\right] \qquad (\delta=0) \qquad (5.87)$$

One concludes that the ratios of the leading terms in $r_{1+}(n,k)$, $r_{L+}(n,k)$, and $r_M(n,k)$ are exactly 1:4:4 if $\delta=0$.

Because comparison of the accuracy is meaningless without considering the operation counts of the schemes being compared, we conclude this section with a comparison of the operation counts of the current and the MacCormack schemes. It is shown in Section 2 and Appendix D that, for each j, *(i)* it requires four multiplications, four additions, and two subtractions for the current scheme to advance one full time step, and *(ii)* it requires five multiplications and four additions for the MacCormack scheme to advance one time step.

6. CONSISTENCY AND THE TRUNCATION ERROR

In Section 5, we studied the question of how well a discrete solution to Eq. (2.54) and (4.1) approximates its analytical counterpart. In this section, we will investigate the circumstances under which an analytical solution may "satisfy" Eq. (2.54). This investigation amounts to a study of the consistency and the truncation error of the current scheme. This study will assume a uniform mesh like that depicted in Fig. 2.1(a). The analysis in this section will be further simplified by assuming $b = 0$, i.e., the mesh is stationary with respect to some coordinate system (x,t). According to a discussion given in Section 3, the last assumption may be made without any loss of generality. Note that consistency and the truncation error are properties to be evaluated at each point within the computational domain. The above uniform-mesh assumption is tantamount to freezing the mesh parameters at their local values. Also, the assumption $b = 0$ is tantamount to introducing a local coordinate system (x,t) such that the mesh is stationary with respect to this system at the local point under consideration.

Before proceeding, we will discuss a general limitation on the ability of an explicit scheme to solve a convection-diffusion problem accurately. As an example, consider Eq. (2.2) (in this section, unless specified otherwise, $\mu > 0$ in Eq. (2.2)) over a domain with $d \geq x \geq 0$ and $t \geq 0$ (see Fig. 6.1). Let the initial data $u(x, 0)$ $(d \geq x \geq 0)$, and the boundary data $u(0,t)$ and $u(d,t)$ $(t > 0)$ be given. Let $u(P_0)$ and $\underline{u}(P_0)$, respectively, denote the values of analytical and discrete solutions at a fixed point P_0. Since a characteristic of Eq. (2.2) is represented by t = constant, the domain of dependence of $u(P_0)$ is the union of $\overline{AB}$, $\overline{BC}$, and $\overline{CD}$. In other words, $u(P_0)$ is dependent on all the initial data, and the boundary data with $t \leq t_0$. Assuming that the discrete solution is generated by an explicit solver, then the domain of dependence of $\underline{u}(P_0)$, contrarily, will include only a subset of the mesh points located on $\overline{AB}$, $\overline{BC}$, and $\overline{CD}$. As an example, consider the MacCormack scheme (see Eq. (D.3)). If the mesh point (j,n) is not on or immediate next to the boundary, then u_j^{n+1} is determined by u_{j-2}^n, u_{j-1}^n, u_j^n, u_{j+1}^n and u_{j+2}^n. As a result, the domain of dependence of $\underline{u}(P_0)$ includes only the mesh points on $\overline{EB}$, $\overline{BC}$, and $\overline{CF}$. Because *(i)* the mesh points that lie on $\overline{AB}$ but not $\overline{EB}$ and those that lie on $\overline{CD}$ but not $\overline{CF}$ do not belong to the domain of dependence of $\underline{u}(P_0)$, and *(ii)* for a *fixed* point P_0, the lengths of $\overline{AE}$ and $\overline{FD}$ are proportional to the ratio $\Delta t/\Delta x$, one may conclude that, *as* $\Delta t, \Delta x \to 0$, *the discrete solution (considered as a function of* Δt *and* Δx*) can not converge to its analytical counterpart unless* $\Delta t/\Delta x \to 0$.

As another example, we consider Eq. (2.2) with $+\infty > x > -\infty$ and $t \geq 0$ (see Fig. 6.2). Let $u(x, 0) = u_0(x)$ where $u_0(x)$ is a given smooth function with period d, i.e., $u_0(x+d) = u_0(x)$. The solution to the above problem also is smooth and has period d in the x-direction. The domain of dependence of $u(P_0)$ is the entire initial line. Assuming that the discrete solution is generated by the MacCormack scheme, then the domain of dependence of $\underline{u}(P_0)$, contrarily, includes only the

mesh points on $\overline{EF}$. If t_0 is small enough, the length of $\overline{EF}$ will be even less than the period d of the initial data. Because the length of $\overline{EF}$ is inversely proportional to the ratio $\Delta t/\Delta x$, one may conclude that, as Δt, $\Delta x \rightarrow 0$, the discrete solution can not converge to its analytical counterpart unless $\Delta t/\Delta x \rightarrow 0$. Note that this conclusion was also mentioned by Fritz John on p.111 in [7].

Contrarily, in the case of pure convection equation Eq. (1.3), the domain of dependence of an analytical solution at (x,t) is a *single* point $(x - at\,,\, 0)$ on the initial line. As a result, $\Delta t/\Delta x \rightarrow 0$ is not required for the convergence of an explicit-scheme discrete solution to its analytical counterpart.

Because an explicit-scheme discrete solution to Eq. (2.2) will not converge to its analytical counterpart as Δt, $\Delta x \rightarrow 0$ without imposing the additional condition $\Delta t/\Delta x \rightarrow 0$, in general one would not expect that, in the limit of Δt, $\Delta x \rightarrow 0$, the nodal values of an analytical solution to Eq. (2.2) will satisfy the explicit scheme which is satisfied by the discrete solution at all Δt and Δx.

In this paper, by definition, a discrete scheme is said to be "strongly consistent" with a PDE when the nodal values of any smooth solution of the PDE satisfy the discrete scheme (i.e., the truncation error = 0) in the limit of Δt, $\Delta x \rightarrow 0$, regardless how the mesh is refined (particularly how $\Delta t/\Delta x$ behaves as Δt, $\Delta x \rightarrow 0$). According to the observation given in the last paragraph, *it should be an exception rather than the norm for an explicit scheme to be strongly consistent with a convection-diffusion equation.* However, perhaps because the strong consistency condition is routinely imposed in the construction of a numerical scheme, there are very few schemes, e.g., the L/D-F scheme, that are not strongly consistent with the PDE to be solved.

At this juncture, note that the term "consistency" as defined on p.44 of [4] represents a concept that involves the numerical scheme, the PDE, and *the rule of mesh refinement* (i.e., how Δt and Δx are related as Δt, $\Delta x \rightarrow 0$). A two-level scheme is said to be consistent with a PDE if the truncation error $\rightarrow 0$ as Δt, $\Delta x \rightarrow 0$ under a given rule of mesh refinement (Note: For multi-level schemes, consistency means more than truncation error $\rightarrow 0$. See p.175 in [4]. However, in this paper, the above definition of consistency for two-level schemes will be extended to a multi-level scheme like the L/D-F scheme. This extended definition should not be confused with the more rigorous definition used in an equivalence theorem given on p.172 of [4].) Generally, one can not say that a scheme is consistent with a PDE without specifying the particular rule of mesh refinement. A scheme may be consistent with a particular PDE under one rule of mesh refinement, and be consistent with another PDE under another rule of refinement. If a scheme is consistent with the same PDE regardless of how Δt, $\Delta x \rightarrow 0$, then it is strongly consistent with this particular PDE.

As will be shown, the current numerical scheme is not a strongly consistent scheme. To expel any misconception that somehow such a scheme is intrinsically inferior than a strongly consistent scheme, next we shall compare the consistency, stability, convergence, and truncation errors of

two model schemes, i.e., the MacCormack scheme, which is strongly consistent with Eq. (2.2), and the L/D-F scheme, which is not.

Let $\tilde{u}(x,t)$ be a smooth function and $\tilde{u}_j^n \overset{def}{=} \tilde{u}(j\Delta x, n\Delta t)$. Then with the aid of Taylor's formula with remainder, it may be shown that

$$[\,\mathrm{FDE(M)}\,]_j^n - [\,\mathrm{PDE}\,]_j^n \equiv [\,\mathrm{ER(M)}\,]_j^n \qquad (\mu \geq 0) \tag{6.1}$$

and

$$[\,\mathrm{FDE(L)}\,]_j^n - [\,\mathrm{PDE}\,]_j^n \equiv [\,\mathrm{ER(L)}\,]_j^n \qquad (\mu \geq 0) \tag{6.2}$$

where

$$[\,\mathrm{PDE}\,]_j^n \overset{def}{=} \left[\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x} - \mu\frac{\partial^2 \tilde{u}}{\partial x^2} \right]_j^n \overset{def}{=} \left[\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x} - \mu\frac{\partial^2 \tilde{u}}{\partial x^2} \right]_{x=j\Delta x,\ t=n\Delta t} \tag{6.3}$$

$$\begin{aligned}[\,\mathrm{FDE(M)}\,]_j^n \overset{def}{=} \frac{1}{\Delta t}\Bigg[& \tilde{u}_j^{n+1} - \frac{\delta}{8}\left(\frac{\delta}{4}+\tau\right)\tilde{u}_{j-2}^n - \frac{1}{2}\left(\frac{\delta}{2}+\tau+\tau^2-\frac{\delta\tau}{2}-\frac{\delta^2}{4}\right)\tilde{u}_{j-1}^n \\ & - \left(1-\frac{\delta}{2}+\frac{3\delta^2}{16}-\tau^2\right)\tilde{u}_j^n - \frac{1}{2}\left(\frac{\delta}{2}-\tau+\tau^2+\frac{\delta\tau}{2}-\frac{\delta^2}{4}\right)\tilde{u}_{j+1}^n \\ & - \frac{\delta}{8}\left(\frac{\delta}{4}-\tau\right)\tilde{u}_{j+2}^n \Bigg]\end{aligned} \tag{6.4}$$

$$[\,\mathrm{ER(M)}\,]_j^n = O(\Delta t) + O[(\Delta x)^2] \tag{6.5}$$

$$\begin{aligned}[\,\mathrm{FDE(L)}\,]_j^n \overset{def}{=} \frac{1}{2\Delta t}\Bigg[& \left(1+\frac{\delta}{2}\right)\tilde{u}_j^{n+1} - \left(\tau+\frac{\delta}{2}\right)\tilde{u}_{j-1}^n \\ & + \left(\tau-\frac{\delta}{2}\right)\tilde{u}_{j+1}^n - \left(1-\frac{\delta}{2}\right)\tilde{u}_j^{n-1} \Bigg]\end{aligned} \tag{6.6}$$

$$[\,\mathrm{ER(L)}\,]_j^n = \mu\left(\frac{\Delta t}{\Delta x}\right)^2\left[\frac{\partial^2 \tilde{u}}{\partial t^2}\right]_j^n + \left(\frac{\Delta t}{\Delta x}\right)^2 O[(\Delta t)^2] + O[(\Delta t)^2] + O[(\Delta x)^2] \tag{6.7}$$

Several comments may be made about Eqs. (6.1) - (6.7):

a.

$$[\,\mathrm{PDE}\,]_j^n = 0 \tag{6.8}$$

if $u = \tilde{u}(x,t)$ is a solution of Eq. (2.2).

b. $u_j^n = \tilde{u}_j^n$ will satisfy the MacCormack scheme Eq. (D.3) if

$$[\,\mathrm{FDE(M)}\,]_j^n = 0 \tag{6.9}$$

As a result, $[\,\mathrm{FDE(M)}\,]_j^n$ may be considered as the approximation of $[\,\mathrm{PDE}\,]_j^n$ associated with the MacCormack scheme. Eq. (6.1) then states that $[\,\mathrm{ER(M)}\,]_j^n$ is the error of this approximation.

c. $u_j^n = \tilde{u}_j^n$ will satisfy the L/D-F scheme Eq. (3.19) if

$$[\,\mathrm{FDE(L)}\,]_j^n = 0 \tag{6.10}$$

As a result, $[\,\mathrm{FDE(L)}\,]_j^n$ may be considered as the approximation of $[\,\mathrm{PDE}\,]_j^n$ associated with the L/D-F scheme. Eq. (6.2) then states that $[\,\mathrm{ER(L)}\,]_j^n$ is the error of this approximation.

d. Let $u = \tilde{u}(x,t)$ be a solution of Eq. (2.2). Then, by definition, $[\,\mathrm{FDE(M)}\,]_j^n$ and $[\,\mathrm{FDE(L)}\,]_j^n$, respectively, are the truncation errors of the MacCormack and the L/D-F schemes at (j,n) [p.20, 4]. Since Eq. (6.8) is satisfied, Eqs. (6.1) and (6.2) imply that

$$[\,\mathrm{FDE(M)}\,]_j^n = [\,\mathrm{ER(M)}\,]_j^n \tag{6.11}$$

and

$$[\,\mathrm{FDE(L)}\,]_j^n = [\,\mathrm{ER(L)}\,]_j^n \tag{6.12}$$

According to Eq. (6.5), $[\,\mathrm{ER(M)}\,]_j^n \to 0$ as $\Delta t,\ \Delta x \to 0$. Thus the truncation error $[\,\mathrm{FDE(M)}\,]_j^n \to 0$ as $\Delta t,\ \Delta x \to 0$. In other words, the MacCormack scheme is *strongly* consistent with Eq. (2.2). On the other hand, according to Eq. (6.7), $[\,\mathrm{ER(L)}\,]_j^n$, and thus the truncation error $[\,\mathrm{FDE(L)}\,]_j^n$, generally does not approach zero as $\Delta t,\ \Delta x \to 0$. As a result, the L/D-F scheme is not strongly consistent with Eq. (2.2). However, $[\,\mathrm{ER(L)}\,]_j^n \to 0$ as $\Delta t,\ \Delta x,\ \Delta t/\Delta x \to 0$. Thus L/D-F scheme is consistent with Eq. (2.2) if the mesh is refined in a way such that $\Delta t/\Delta x \to 0$ as $\Delta t,\ \Delta x \to 0$.

e. Because *(i)* a discrete solution to Eq. (2.2) generated by an explicit scheme like the MacCormack scheme can not converge to its analytical counterpart without imposing the condition that $\Delta t/\Delta x \to 0$ as $\Delta t,\ \Delta x \to 0$, *(ii)* MacCormack scheme is strongly consistent with Eq. (2.2), and *(iii)* Lax's equivalence theorem [p.45, 4] states that, given a properly posed initial-value problem and a finite-difference approximation to it that satisfies the consistency condition, stability is the necessary and sufficient condition for convergence, one comes to the conclusion that, *in solving a properly posed convection-diffusion problem like that depicted in Fig. 6.2, stability of the MacCormack scheme will require that* $\Delta t/\Delta x \to 0$ *as* $\Delta t,\ \Delta x \to 0$. Note that the term "stability" referred to in Lax's equivalence theorem has a meaning different from that defined in Section 4.

f. As will be shown in Appendix E, the Lax stability of the MacCormack scheme requires that the mesh be refined in a way such that the parameter δ remains bounded as $\Delta t, \Delta x \to 0$. In the case where $\mu > 0$, this implies that

$$\frac{\Delta t}{\Delta x} = O(\Delta x) \tag{6.13}$$

It follows from Eq. (6.13) that $\Delta t/\Delta x \to 0$ as $\Delta t, \Delta x \to 0$. Assuming that $u = \tilde{u}(x,t)$ satisfies Eq. (2.2), then Eqs. (6.5), (6.7), and (6.11) - (6.13) imply that

$$[\,\mathrm{FDE(M)}\,]_j^n = O[\,(\Delta x)^2\,] \quad \text{and} \quad [\,\mathrm{FDE(L)}\,]_j^n = O[\,(\Delta x)^2\,] \tag{6.14}$$

Thus the MacCormack scheme has no advantage over the L/D-F scheme in the order of the truncation error if the Lax stability is considered.

This completes the comparisons between the MacCormack and the L/D-F schemes. It has been shown that a scheme that is strongly consistent may not have an intrinsic advantage over a scheme that is not.

Next we shall study the consistency and the truncation error of the current scheme, particularly the two-level, two-dependent-variable discrete equations Eqs. (2.56) and (2.57). It will be shown that these two equations are consistent with a pair of PDEs if $\Delta t/\Delta x \to 0$ as $\Delta t, \Delta x \to 0$. One of these PDE's is Eq. (2.2).

To proceed, let $\tilde{u}(x,t)$ and $\tilde{v}(x,t)$ be smooth functions. Let

$$\tilde{w} \stackrel{def}{=} \tilde{v} - \frac{\partial \tilde{u}}{\partial x} \tag{6.15}$$

Let $\tilde{u}_j^n \stackrel{def}{=} \tilde{u}(j\Delta x, n\Delta t)$, $\tilde{v}_j^n \stackrel{def}{=} \tilde{v}(j\Delta x, n\Delta t)$ and $\tilde{w}_j^n \stackrel{def}{=} \tilde{w}(j\Delta x, n\Delta t)$. Also let

$$\begin{aligned}
[F1]_j^n \stackrel{def}{=}\ & \tilde{u}_j^{n+1} - \frac{1}{2}\left[\tau + \frac{\delta(1-\tau)}{1-\tau^2+\delta}\right]\left[(1+\tau)\tilde{u}_{j-1}^n + \frac{1-\tau^2-\delta}{4}\Delta x\,\tilde{v}_{j-1}^n\right] \\
& - \frac{1-\tau^2}{1-\tau^2+\delta}\left[(1-\tau^2)\tilde{u}_j^n - \frac{\tau(1-\tau^2-\delta)}{4}\Delta x\,\tilde{v}_j^n\right] \\
& + \frac{1}{2}\left[\tau - \frac{\delta(1+\tau)}{1-\tau^2+\delta}\right]\left[(1-\tau)\tilde{u}_{j+1}^n - \frac{1-\tau^2-\delta}{4}\Delta x\,\tilde{v}_{j+1}^n\right]
\end{aligned} \tag{6.16}$$

$$[F2]_j^n \stackrel{def}{=} \Delta x\,\tilde{v}_j^{n+1} + \frac{1}{2}\left[\tau + \frac{\delta(1-\tau)}{1-\tau^2+\delta}\right]\left[4\frac{1-\tau^2}{1-\tau^2+\delta}\tilde{u}_{j-1}^n + (1-\tau)\frac{1-\tau^2-\delta}{1-\tau^2+\delta}\Delta x\,\tilde{v}_{j-1}^n\right]$$

$$- \left[\frac{1-\tau^2}{1-\tau^2+\delta} \right]^2 \left[4\tau \tilde{u}_j^n + (1-\tau^2-\delta)\Delta x \tilde{v}_j^n \right]$$

$$+ \frac{1}{2} \left[\tau - \frac{\delta(1+\tau)}{1-\tau^2+\delta} \right] \left[4 \frac{1-\tau^2}{1-\tau^2+\delta} \tilde{u}_{j+1}^n - (1+\tau) \frac{1-\tau^2-\delta}{1-\tau^2+\delta} \Delta x \tilde{v}_{j+1}^n \right] \tag{6.17}$$

Then, with the aid of Taylor's formula with remainder and the assumptions that

$$\delta > 0 \quad \text{and} \quad 1 > \tau^2 \tag{6.18}$$

It may be shown that

$$[\text{FDE1}]_j^n - [\text{PDE}]_j^n \equiv [\text{ER1}]_j^n \tag{6.19}$$

and

$$[\text{FDE2}]_j^n - \tilde{w}_j^n \equiv [\text{ER2}]_j^n \tag{6.20}$$

Here $[\text{PDE}]_j^n$ is defined in Eq. (6.3). Also

$$[\text{FDE1}]_j^n \stackrel{def}{=} \frac{1}{\Delta t} [F1]_j^n \tag{6.21}$$

$$[\text{FDE2}]_j^n \stackrel{def}{=} \frac{(1-\tau^2+\delta)^2}{4\delta(1-\tau^2)\Delta x} [F2]_j^n \tag{6.22}$$

$$[\text{ER1}]_j^n \stackrel{def}{=} \mu \frac{1-\tau^2-\delta}{1-\tau^2+\delta} \left[\frac{\partial \tilde{w}}{\partial x} \right]_j^n + \frac{1}{2} \left[\left[\frac{\partial^2 \tilde{u}}{\partial t^2} \right]_j^n - a^2 \left[\frac{\partial^2 \tilde{u}}{\partial x^2} \right]_j^n \right] \Delta t$$

$$+ \frac{a}{6} \left\{ \left[\frac{\partial^3 \tilde{u}}{\partial x^3} \right]_j^n (\Delta x)^2 - \frac{3}{4} \left[\frac{\partial^2 \tilde{v}}{\partial x^2} \right]_j^n \frac{(1-\tau^2-\delta)[(\Delta x)^2 - a^2(\Delta t)^2]}{1-\tau^2+\delta} \right\} + O[(\Delta t)^2]$$

$$+ \left[\tau + \frac{\delta(1-\tau)}{1-\tau^2+\delta} \right] \frac{1}{\Delta t} \left\{ (1+\tau) \times O[(\Delta x)^4] + (1-\tau^2-\delta) \times O[(\Delta x)^4] \right\}$$

$$+ \left[\tau - \frac{\delta(1+\tau)}{1-\tau^2+\delta} \right] \frac{1}{\Delta t} \left\{ (1-\tau) \times O[(\Delta x)^4] + (1-\tau^2-\delta) \times O[(\Delta x)^4] \right\} \tag{6.23}$$

$$[\mathrm{ER2}]_j^n \stackrel{def}{=} \frac{(1-\tau^2+\delta)^2(\Delta x)^2}{16\mu(1-\tau^2)}\left[\frac{\partial \tilde{v}}{\partial t}\right]_j^n - \frac{a(1-\tau^2-\delta)^2(\Delta x)^2}{16\mu(1-\tau^2)}\left[\frac{\partial \tilde{v}}{\partial x}\right]_j^n$$

$$+\frac{a(1-\tau^2)(\Delta x)^2}{8\mu}\left[\frac{\partial^2 \tilde{u}}{\partial x^2}\right]_j^n + \frac{(1-\tau^2+\delta)^2}{4\delta(1-\tau^2)} \times O[(\Delta t)^2]$$

$$+\left[1+\frac{\tau}{\delta}(1-\tau^2)\right] \times \left\{O[(\Delta x)^2] + \left[\frac{1-\tau^2-\delta}{1+\tau}\right] \times O[(\Delta x)^2]\right\}$$

$$+\left[1-\frac{\tau}{\delta}(1-\tau^2)\right] \times \left\{O[(\Delta x)^2] + \left[\frac{1-\tau^2-\delta}{1-\tau}\right] \times O[(\Delta x)^2]\right\} \tag{6.24}$$

The significance of Eqs. (6.19) - (6.24) will be discussed in the following comments:

a. $[\mathrm{FDE1}]_j^n$ and $[\mathrm{FDE2}]_j^n$, respectively, may be considered approximations of $[\mathrm{PDE}]_j^n$ and $\tilde{w}_j^n$. Eqs. (6.19) and (6.20) then state that $[\mathrm{ER1}]_j^n$ and $[\mathrm{ER2}]_j^n$, respectively, are the errors of these approximations.

b. Eqs. (2.56) and (2.57), respectively, are equivalent to

$$[\mathrm{FDE1}]_j^n = 0 \tag{6.25}$$

and

$$[\mathrm{FDE2}]_j^n = 0 \tag{6.26}$$

if $\tilde{u}_j^n$ and $\tilde{v}_j^n$ in Eqs. (6.25) and (6.26), respectively, are replaced by γ_j^n and α_j^n.

c. With the aid of the observations made in (a) and (b), and Eqs. (6.3) and (6.15), one concludes that Eqs. (2.56) and (2.57), respectively, may be considered as the discrete approximations of Eq. (2.2) and the PDE

$$v - \frac{\partial u}{\partial x} = 0 \tag{6.27}$$

with the understanding that γ_j^n and α_j^n, respectively, are the discrete counterparts of u and v.

d. Let $u = \tilde{u}(x,t)$ and $v = \tilde{v}(x,t)$ be a solution of the system of PDEs Eqs. (2.2) and (6.27). Then, by definition, $[\mathrm{FDE1}]_j^n$ and $[\mathrm{FDE2}]_j^n$ are the truncation errors of the discrete equations Eqs. (2.56) and (2.57). Furthermore, we have $[\mathrm{PDE}]_j^n = 0$ and $w_j^n = 0$. Thus

$$[\mathrm{FDE1}]_j^n = [\mathrm{ER1}]_j^n \quad \text{and} \quad [\mathrm{FDE2}]_j^n = [\mathrm{ER2}]_j^n \tag{6.28}$$

i.e., $[\mathrm{ER1}]_j^n$ and $[\mathrm{ER2}]_j^n$ become the truncation errors.

e. Since μ and a are constants, parameters τ and δ vary as Δt and Δx vary. Generally, τ and δ do not approach certain limits as $\Delta t, \Delta x \to 0$. In spite of this and the fact that $[\mathrm{ER1}]_j^n$ and $[\mathrm{ER2}]_j^n$ are dependent on τ and δ, it is shown in Appendix E that

$$\lim_{\Delta t\,,\,\Delta x \to 0} [\mathrm{ER1}]_j^n = 0 \tag{6.29}$$

and

$$\lim_{\Delta t/\Delta x\,,\,\Delta x \to 0} [\mathrm{ER2}]_j^n = 0 \tag{6.30}$$

if $u = \tilde{u}(x,t)$ and $v = \tilde{v}(x,t)$ satisfy Eq. (6.27). Note that the notation $\Delta t \to 0$ does not appear on the left side of Eq. (6.30) because $\Delta t \to 0$ if $\Delta t/\Delta x$, $\Delta x \to 0$. Furthermore, it is shown in Appendix E that

$$[\mathrm{ER1}]_j^n = O[(\Delta x)^2] \quad \text{and} \quad [\mathrm{ER2}]_j^n = O[(\Delta x)^2] \tag{6.31}$$

if $u = \tilde{u}(x,t)$ and $v = \tilde{v}(x,t)$ satisfy Eq. (6.27) and the rule of refinement is such that δ remains bounded as $\Delta t, \Delta x \to 0$.

f. Combining (a) - (e), one may conclude that Eqs. (2.56) and (2.57) are consistent with Eqs. (2.2) and (6.27) if the rule of mesh refinement is such that $\Delta t/\Delta x \to 0$ as $\Delta t, \Delta x \to 0$.

g. Because of Eq. (6.27), convergence of a discrete solution (γ_j^n, α_j^n) to its analytical counterpart (u, v) implies that $\gamma_j^n \to u$ and $\alpha_j^n \to \partial u/\partial x$. This is consistent with the interpretation of γ_j^n and α_j^n given in Eqs. (1.17a) and (1.17b).

Finally we consider the special case in which $\delta = 0$ (i.e., $\mu = 0$). For this case, it may be shown that

$$\frac{1}{\Delta t}[F1]_j^n - \left[\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x}\right]_j^n \equiv \frac{\Delta t}{2}\left[\frac{\partial^2 \tilde{u}}{\partial t^2} - a^2\frac{\partial^2 \tilde{u}}{\partial x^2}\right]_j^n + O[(\Delta x)^2] + O[(\Delta t)^2] \tag{6.32}$$

$$\frac{1}{\Delta t\,\Delta x}[F2]_j^n - \left[\frac{\partial \tilde{w}}{\partial t} - a\frac{\partial \tilde{w}}{\partial x} + \frac{\partial}{\partial x}\left[\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x}\right]\right]_j^n$$

$$\equiv \frac{\Delta t}{2}\left[\frac{\partial^2 \tilde{w}}{\partial t^2} - a^2\frac{\partial^2 \tilde{w}}{\partial x^2} + \frac{\partial}{\partial x}\left[\frac{\partial^2 \tilde{u}}{\partial t^2} - a^2\frac{\partial^2 \tilde{u}}{\partial x^2}\right]\right]_j^n + O[(\Delta x)^2] + O[(\Delta t)^2] \tag{6.33}$$

Using Eqs. (6.32) and (6.33) and some comments given previously, one concludes that Eqs. (2.56) and (2.57) are strongly consistent with Eq. (1.3) and

$$\frac{\partial w}{\partial t} - a\frac{\partial w}{\partial x} + \frac{\partial}{\partial x}\left[\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}\right] = 0 \tag{6.34}$$

Eqs. (1.3) and (6.34) are equivalent to Eq. (1.3) and

$$\frac{\partial w}{\partial t} - a\frac{\partial w}{\partial x} = 0 \tag{6.35}$$

Eqs. (1.3) and (6.35) are similar in their forms. They, respectively, represent the wave motions propagating with the speeds $+a$ and $-a$.

Let $u = \tilde{u}(x,t)$ and $v = \tilde{v}(x,t)$ satisfy Eqs. (1.3) and (6.35). Then the lowest-order terms on the right sides of Eqs. (6.32) and (6.33) vanish. Thus Eqs. (2.56) and (2.57) are strongly consistent with Eqs. (1.3) and (6.35) with their truncation errors being $O[(\Delta x)^2] + O[(\Delta t)^2]$.

7. NUMERICAL EVALUATION

In this section, the current method will be compared numerically with the L/D-F method and the MacCormack method. During this comparison, many theoretical results developed previously will also be evaluated using the numerical results presented.

To simplify our effort, the numerical problems to be considered have the common initial condition

$$u_j^0 = \sin(2\pi j\Delta x) \quad , \qquad j = 0, \pm 1, \pm 2, \cdots \tag{7.1}$$

where

$$\Delta x = \frac{1}{K} \tag{7.2}$$

and $K \geq 3$ is the integer introduced in Eq. (4.1). According to Eqs. (7.1) and (7.2), u_j^0 is periodic over every unit length and every K mesh intervals in the x-direction. In this section, we shall continue to use the mesh depicted in Fig. 2.1(a) and assume the periodic condition, i.e., $u_j^n = u_{j+K}^n$, $j = 0, \pm 1, \pm 2, \cdots, n = 0, 1, 2, \cdots$.

The analytical problem corresponding to the above numerical problem will be specified by the periodic condition $u(x+K\Delta x,t) = u(x,t)$ and the initial condition $u(x, 0) = I(x)$ where $I(x)$ is determined according to Eqs. (5.7) and (5.8). With the aid of Eqs. (4.7), (5.7), (5.8), (7.1), and (7.2), it may be shown that

$$b_0 = 0 \quad , \quad b_1 = \frac{1}{2i} \quad , \quad b_2 = b_3 = \cdots = b_{K-2} = 0 \quad , \quad b_{K-1} = -\frac{1}{2i} \tag{7.3}$$

$$\theta_1 = \frac{2\pi}{K} \quad , \quad \theta_{K-1} = -\frac{2\pi}{K} \quad , \tag{7.4}$$

and

$$I(x) = \sin(2\pi x) \tag{7.5}$$

By using Eqs. (5.15) and (7.2) - (7.4), one obtains the analytical solution to Eq. (2.2), i.e.,

$$u = u_a(x,t) \stackrel{def}{=} e^{-4\pi^2\mu t} \sin[2\pi(x-at)] \tag{7.6}$$

Note that parameters γ_j^0 and α_j^0, which are needed in the initiation of the current numerical procedure, may also be evaluated by using Eqs. (5.1), (5.4), (7.1), and (7.5).

Combining Eqs. (5.17), (5.52), (5.53), and (7.2) - (7.4), it may be shown that

$$u_a(x_j^n,t^n) = e^{-4\pi^2\mu n\Delta t} \sin\{2\pi[j\Delta x - n(a-b)\Delta t]\} = u_a(j,n,1) + u_a(j,n,K-1) \tag{7.7}$$

with

$$u_a(j,n,1) = \frac{1}{2i} D(n\Delta t)\, e^{i\phi_j^n} \tag{7.8}$$

and

$$u_a(j,n,K-1) = -\frac{1}{2i} D(n\Delta t)\, e^{-i\phi_j^n} \tag{7.9}$$

Here

$$D(t) \stackrel{def}{=} e^{-4\pi^2\mu t} \qquad (t \geq 0) \tag{7.10}$$

is the decay factor of $u_a(x,t)$ and

$$\phi_j^n \stackrel{def}{=} 2\pi(j - n\tau)/K \tag{7.11}$$

is a phase angle. $u_a(x_j^n,t^n)$ is the value of the analytical solution at the mesh point (j,n). Its numerical counterpart in the present method is γ_j^n. By using Eqs. (2.23), (2.24), (5.13), (5.14), (5.54), and (7.3), one concludes that

$$\gamma_j^n = \underline{u}(j,n,1) + \underline{u}(j,n,K-1) \tag{7.12}$$

Let

$$R_1(j,n) \stackrel{def}{=} \frac{\gamma_j^n - u_a(x_j^n,t^n)}{D(n\Delta t)} \tag{7.13}$$

In other words, $R_1(j,n)$ *is the error of the numerical solution* γ_j^n *normalized by the decay factor of* $u_a(x_j^n,t^n)$. Eqs. (5.55) and (7.7) - (7.13) imply that

$$R_1(j,n) = \frac{1}{2i}\left[r_1(n,1)\, e^{i\phi_j^n} - r_1(n,K-1)\, e^{-i\phi_j^n} \right] \tag{7.14}$$

Substituting Eqs. (5.46), (5.48), and (5.49) into Eq. (7.14) and using Eq. (7.4) one has

$$\begin{aligned} R_1(j,n) \doteq\ & n\left\{ \frac{\tau\Delta_1(\tau,\delta)}{24}(2\pi/K)^3 \cos(\phi_j^n) \right. \\ & \left. + \frac{\delta}{192}\left[\frac{1-5\tau^2}{1-\tau^2}\Delta_1(\tau,\delta) - 4\tau^2 \right](2\pi/K)^4 \sin(\phi_j^n) + O[(2\pi/K)^5] \right\} \\ & - \frac{(1-\tau^2)^2 - \delta^2}{32(1-\tau^2)}\left\{ \frac{\tau\delta}{1-\tau^2}(2\pi/K)^3 \cos(\phi_j^n) \right. \\ & \left. + \frac{1}{24}\left[(1+3\tau^2) + \frac{3\delta^2(1-9\tau^2)}{(1-\tau^2)^2} \right](2\pi/K)^4 \sin(\phi_j^n) + O[(2\pi/K)^5] \right\} \end{aligned} \tag{7.15}$$

According to Eq. (5.56), $u_{ax}(x_j^n,t^n)$ is the value of $\partial u_a(x,t)/\partial x$ at the mesh point (j,n). Its numerical counterpart in the current method is α_j^n. By using Eqs. (5.17), (5.56), (5.57), and (7.2) - (7.4), one concludes that

$$u_{ax}(x_j^n,t^n) = 2\pi e^{-4\pi^2\mu n\Delta t}\cos\{2\pi[j\Delta x - n(a-b)\Delta t]\} = u_{ax}(j,n,1) + u_{ax}(j,n,K-1) \tag{7.16}$$

with

$$u_{ax}(j,n,1) = \pi D(n\Delta t)e^{i\phi_j^n} \tag{7.17}$$

$$u_{ax}(j,n,K-1) = \pi D(n\Delta t)e^{-i\phi_j^n} \tag{7.18}$$

Moreover, Eqs. (2.23), (2.24), (5.13), (5.14), (5.58), and (7.3) imply that

$$\alpha_j^n = \underline{u}_x(j,n,1) + \underline{u}_x(j,n,K-1) \tag{7.19}$$

Let

$$R_2(j,n) \stackrel{def}{=} \frac{\alpha_j^n - u_{ax}(x_j^n,t^n)}{2\pi D(n\Delta t)} \tag{7.20}$$

i.e., $R_2(j,n)$ *is the error of* α_j^n *normalized by the "amplitude" of* $u_{ax}(x_j^n,t^n)$. Then Eqs. (5.59) and (7.16) - (7.20) may be used to show that

$$R_2(j,n) = \frac{1}{2}\left[r_2(n,1)e^{i\phi_j^n} + r_2(n,K-1)e^{-i\phi_j^n}\right] \tag{7.21}$$

Substituting Eqs. (5.46), (5.50), and (5.51) into Eq. (7.21) and using Eq. (7.4), one has

$$\begin{aligned} R_2(j,n) \doteq\ & n\left\{-\frac{\tau\Delta_1(\tau,\delta)}{24}(2\pi/K)^3\sin(\phi_j^n)\right. \\ & \left. + \frac{\delta}{192}\left[\frac{1-9\tau^2}{1-\tau^2}\Delta_1(\tau,\delta) - 4\tau^2\right](2\pi/K)^4\cos(\phi_j^n) + O[(2\pi/K)^5]\right\} \\ & - \frac{\tau\delta}{2(1-\tau^2)}(2\pi/K)\sin(\phi_j^n) \\ & + \frac{1}{48}\left[(1+3\tau^2) + \frac{3\delta^2(1-5\tau^2)}{(1-\tau^2)^2}\right](2\pi/K)^2\cos(\phi_j^n) + O[(2\pi/K)^3] \end{aligned} \tag{7.22}$$

According to Eqs. (7.3), (B.31), and (B.33), the numerical counterpart of $u_a(x_j^n,t^n)$ in the L/D-F method is

$$\underline{u}_L(j,n) = \underline{u}_L(j,n,1) + \underline{u}_L(j,n,K-1) \tag{7.23}$$

Let

$$R_L(j,n) \stackrel{def}{=} \frac{\underline{u}_L(j,n) - u_a(x_j^n, t^n)}{D(n\Delta t)} \tag{7.24}$$

Then Eqs. (5.60), (7.7) - (7.9), and (7.23) imply that

$$R_L(j,n) = \frac{1}{2i}\left[r_L(n,1)\, e^{i\phi_j^n} - r_L(n,K-1)\, e^{-i\phi_j^n} \right] \tag{7.25}$$

Substituting Eqs. (5.61) - (5.63) into Eq. (7.25) and using Eq. (7.4), one has

$$\begin{aligned} R_L(j,n) \doteq\ & n\left\{ \frac{\tau^2\delta}{4}(2\pi/K)^2 \sin(\phi_j^n) + \frac{\tau}{6}(1-\tau^2)(1-\frac{3}{4}\delta^2)(2\pi/K)^3 \cos(\phi_j^n) \right. \\ & + \frac{\delta}{96}\left[9\delta^2\tau^2(1-\tau^2) + 3\delta\tau^4 + 4\tau^2 - \frac{3}{2}\delta^2 + 2 \right](2\pi/K)^4 \sin(\phi_j^n) + O[(2\pi/K)^5] \Big\} \\ & + \frac{1}{4}(1-\frac{\delta^2}{4})\left\{ \tau^2(2\pi/K)^2 \sin(\phi_j^n) - \frac{\tau\delta}{2}(1-2\tau^2)(2\pi/K)^3 \cos(\phi_j^n) \right. \\ & - \frac{1}{48}\left[4\tau^2(4-9\tau^2) + 3\delta^2(15\tau^4 - 12\tau^2 + 1) \right](2\pi/K)^4 \sin(\phi_j^n) \\ & \left. + O[(2\pi/K)^5] \right\} \end{aligned} \tag{7.26}$$

According to Eqs. (7.3), (D.5), and (D.6), the numerical counterpart of $u_a(x_j^n, t^n)$ in the MacCormack scheme is

$$\underline{u}_M(j,n) = \underline{u}_M(j,n,1) + \underline{u}_M(j,n,K-1) \tag{7.27}$$

Let

$$R_M(j,n) \stackrel{def}{=} \frac{\underline{u}_M(j,n) - u_a(x_j^n, t^n)}{D(n\Delta t)} \tag{7.28}$$

Then Eqs. (5.64), (7.7) - (7.9), and (7.28) imply that

$$R_M(j,n) = \frac{1}{2i}\left[r_M(n,1)\, e^{i\phi_j^n} - r_M(n,K-1)\, e^{-i\phi_j^n} \right] \tag{7.29}$$

Substituting Eq. (5.65) into Eq. (7.29) and using Eq. (7.4), one has

$$R_M(j,n) \doteq n\left\{ \frac{\tau(1-\tau^2)}{6}(2\pi/K)^3 \cos(\phi_j^n) \right.$$

$$+\frac{1}{48}\left[\delta(1-6\tau^2)-6\tau^2(1-\tau^2)\right](2\pi/K)^4\sin(\phi_j^n)+O[(2\pi/K)^5]\Big\} \qquad (7.30)$$

Let $\tau = 0$. Then Eqs. (7.15), (7.22), (7.26), and (7.30), respectively, are reduced to

$$R_1(j,n) \doteq n\left\{\frac{\delta(1-3\delta^2)}{192}(2\pi/K)^4\sin(\phi_j^n)+O[(2\pi/K)^5]\right\}$$

$$-\frac{(1-\delta^2)(1+3\delta^2)}{768}(2\pi/K)^4\sin(\phi_j^n)+O[(2\pi/K)^5] \qquad (\tau=0) \qquad (7.31)$$

$$R_2(j,n) \doteq n\left\{\frac{\delta(1-3\delta^2)}{192}(2\pi/K)^4\cos(\phi_j^n)+O[(2\pi/K)^5]\right\}$$

$$+\frac{(1+3\delta^2)}{48}(2\pi/K)^2\cos(\phi_j^n)+O[(2\pi/K)^3] \qquad (\tau=0) \qquad (7.32)$$

$$R_L(j,n) \doteq n\left\{\frac{\delta(4-3\delta^2)}{192}(2\pi/K)^4\sin(\phi_j^n)+O[(2\pi/K)^5]\right\}$$

$$-\frac{\delta^2}{64}\left(1-\frac{\delta^2}{4}\right)(2\pi/K)^4\sin(\phi_j^n)+O[(2\pi/K)^5] \qquad (\tau=0) \qquad (7.33)$$

and

$$R_M(j,n) \doteq n\left\{\frac{\delta}{48}(2\pi/K)^4\sin(\phi_j^n)+O[(2\pi/K)^5]\right\} \qquad (\tau=0) \qquad (7.34)$$

Let τ and δ satisfy the optimal condition Eq. (5.70), then Eqs. (7.15) and (7.22), respectively, are reduced to

$$R_1(j,n) \doteq n\left\{-\frac{\tau^2(1-\tau^2)}{48\sqrt{3}}(2\pi/K)^4\sin(\phi_j^n)+O[(2\pi/K)^5]\right\}$$

$$-\frac{1-\tau^2}{48}\left\{\frac{\tau}{\sqrt{3}}(2\pi/K)^3\cos(\phi_j^n)+\frac{1}{12}(1-3\tau^2)(2\pi/K)^4\sin(\phi_j^n)\right.$$

$$\left.+O[(2\pi/K)^5]\right\} \qquad (1-\tau^2=\sqrt{3}\,\delta) \qquad (7.35)$$

and

$$R_2(j,n) \doteq n\left\{-\frac{\tau^2(1-\tau^2)}{48\sqrt{3}}(2\pi/K)^4\cos(\phi_j^n)+O[(2\pi/K)^5]\right\}$$
$$-\frac{\tau}{2\sqrt{3}}(2\pi/K)\sin(\phi_j^n)+\frac{1}{24}(1-\tau^2)(2\pi/K)^2\cos(\phi_j^n)$$

$$+O[(2\pi/K)^3] \qquad (1-\tau^2=\sqrt{3}\,\delta) \tag{7.36}$$

According to Eqs. (5.46), (5.48), and (5.49), $r_1(n,k)$ may be approximated by the sum of the expressions on the right sides of Eqs. (5.48) and (5.49). This sum may be converted to the approximate form of $R_1(j,n)$ given in Eq. (7.15) if one carries out the following substitution:

$$(\theta_k)^l \rightarrow \left\{\begin{array}{ll} (2\pi/K)^l\sin(\phi_j^n) & \text{if } l \text{ is even} \\ \\ \dfrac{1}{i}(2\pi/K)^l\cos(\phi_j^n) & \text{if } l \text{ is odd} \end{array}\right\} \tag{7.37}$$

The same relation also exists between $r_L(n,k)$ and $R_L(j,n)$, and between $r_M(n,k)$ and $R_M(j,n)$. A similar relation also exists between $r_2(n,k)$ and $R_2(j,n)$. However, for this case, the rule of substitution is

$$(\theta_k)^l \rightarrow \left\{\begin{array}{ll} (2\pi/K)^l\cos(\phi_j^n) & \text{if } l \text{ is even} \\ \\ i(2\pi/K)^l\sin(\phi_j^n) & \text{if } l \text{ is odd} \end{array}\right\} \tag{7.38}$$

The approximations for the spurious parts $r_{1-}(n,k)$, $r_{2-}(n,k)$, and $r_{L-}(n,k)$, respectively, given by Eqs. (5.49), (5.51), and (5.63) are used in the above derivation of Eqs. (7.15), (7.22), and (7.26). In the case where $\delta = 0$ or δ is very close to zero, these spurious-part approximations may no longer be valid. Fortunately, the spurious part generally is negligible compared with the principal part in a calculation with large n. Thus we may completely omit the spurious parts in Eqs. (7.15), (7.22), and (7.26) if δ is very close to zero and n is large.

Subject to the modifications required by the substitution rule given in Eqs. (7.37) and (7.38), the comments made in Section 5 about the approximate forms of $r_1(n,k)$, $r_2(n,k)$, $r_L(n,k)$, and $r_M(n,k)$ are applicable to those of $R_1(j,n)$, $R_2(j,n)$, $R_L(j,n)$, and $R_M(j,n)$. Particularly, the leading terms in the principal parts of $R_1(j,n)$ and $R_2(j,n)$ also vanish if $\Delta_1(\tau,\delta) = 0$. Because the factor $2\pi/K \rightarrow 0$ as $K \rightarrow +\infty$ (i.e., $\Delta x \rightarrow 0$), the leading term becomes more dominant as $K \rightarrow +\infty$. As a result, the effect of the leading-term annihilation on the accuracy of a numerical method becomes more pronounced as $K \rightarrow +\infty$.

Eq. (7.6) implies that

$$a\frac{\partial u_a(x,t)}{\partial x} = 2\pi a\, e^{-4\pi^2\mu t}\cos[2\pi(x-at)] \tag{7.39}$$

and

$$\mu\frac{\partial^2 u_a(x,t)}{\partial x^2} = -4\pi^2\mu\, e^{-4\pi^2\mu t}\sin[2\pi(x-at)] \tag{7.40}$$

Thus, in the case where $u = u_a(x,t)$, the relative importance of the convection and the diffusion terms in Eq. (2.2) may be determined by comparing the "amplitude" of the expressions on the right sides of Eqs. (7.39) and (7.40). As a result, in the following numerical study, the solution $u = u_a(x,t)$ will be referred to as: *(i)* convection dominant if $|a| \gg 2\pi\mu$, *(ii)* convection-diffusion comparable if $|a| \approx 2\pi\mu$, and *(iii)* diffusion dominant if $|a| \ll 2\pi\mu$.

In the following numerical study, each test problem is defined by the values of *(i)* the physical parameters a and μ, and *(ii)* the mesh parameters b, n, Δt, and K $(= \frac{1}{\Delta x})$. After n time steps, the numerical error of a test problem will be measured by

$$\bar{E}(b,n,\Delta t,K) \stackrel{def}{=} \log_{10}\left[\frac{1}{D(n\Delta t)}\frac{1}{K}\sum_{j=0}^{K-1}|u_j^n - u_a(x_j^n,t^n)|\right] \tag{7.41}$$

where u_j^n is the numerical solution at the mesh point (j,n). Roughly speaking, the negative of $\bar{E}(b,n,\Delta t,K)$ represents the average number of correct significant figures in u_j^n, $j = 0, 1, 2, \cdots, K-1$. Note that, in the current paper, we will not distinguish between the exact solution of a numerical scheme and the actual numerical solution of the same scheme, i.e., the roundoff error is assumed to be negligible.

With the aid of Eqs. (7.13), (7.24), and (7.28), one concludes that

$$\bar{E}(b,n,\Delta t,K) = \left\{\begin{array}{ll} \log_{10}\bar{R}_1(b,n,\Delta t,K) & (\text{ Present scheme }) \\ \log_{10}\bar{R}_L(b,n,\Delta t,K) & (\text{ L/D-F scheme }) \\ \log_{10}\bar{R}_M(b,n,\Delta t,K) & (\text{ MacCormack scheme }) \end{array}\right\} \tag{7.42}$$

Here

$$\bar{R}_1(b,n,\Delta t,K) \stackrel{def}{=} \frac{1}{K}\sum_{j=0}^{K-1}|R_1(j,n)| \tag{7.43}$$

$$\bar{R}_L(b,n,\Delta t,K) \stackrel{def}{=} \frac{1}{K}\sum_{j=0}^{K-1}|R_L(j,n)| \tag{7.44}$$

and

$$\bar{R}_M(b,n,\Delta t,K) \stackrel{def}{=} \frac{1}{K}\sum_{j=0}^{K-1} |R_M(j,n)| \tag{7.45}$$

, respectively, are the averages of $|R_1(j,n)|$, $|R_L(j,n)|$ and $|R_M(j,n)|$ at time level n. Note that $R_1(j,n)$, $R_L(j,n)$, and $R_M(j,n)$, implicitly, are functions of b, Δt, and K. In our future discussions, $\bar{E}(b,n,\Delta t,K)$, $\bar{R}_1(b,n,\Delta t,K)$, $\bar{R}_L(b,n,\Delta t,K)$, and $\bar{R}_M(b,n,\Delta t,K)$, respectively, may be abbreviated as $\bar{E}$, $\bar{R}_1$, $\bar{R}_L$, and $\bar{R}_M$.

To pave the way for the interpretation of the numerical results to be presented, several approximations of $\bar{R}_1$, $\bar{R}_L$, and $\bar{R}_M$ will be introduced in the following discussions:

a. Let $1-\tau^2 = \sqrt{3}\,\delta$. Then the approximation of $R_1(j,n)$ given in Eq. (7.35) may be applicable. Let $R_1(j,n)$ be dominant by the leading term in its principal part. Then it may be shown that

$$\bar{R}_1 \doteq \frac{2n\pi^3\tau^2|1-\tau^2|}{3\sqrt{3}K^4} \qquad (1-\tau^2 = \sqrt{3}\,\delta \ , \quad n, K \gg 1) \tag{7.46}$$

Note that, in obtaining Eq. (7.46), we use the identity

$$\lim_{K\to+\infty}\left\{\frac{1}{K}\sum_{j=0}^{K-1}|\sin(\phi_j^n + C)|\right\} = \frac{2}{\pi} \tag{7.47}$$

where C is any real number which is independent of j.

The parameters Re, τ_0, δ_0, and $(\Delta t)_0$ were defined in terms of $(a-b)$, μ, and Δx $(=1/K)$ in Section 5. By these definitions, we have $1-\tau_0^2 = \sqrt{3}\,\delta_0$. Furthermore, Eqs. (5.73) - (5.75) imply that

$$|\tau_0| \ll 1 \quad \text{and} \quad \delta_0 \doteq \frac{1}{\sqrt{3}} \qquad \text{if } |\mathrm{Re}| \ll 1 \tag{7.48}$$

Combining Eqs. (2.17), (2.18), (7.46), and (7.48), one has

$$\bar{R}_1(b,n,(\Delta t)_0,K) \doteq \frac{\pi^3(a-b)^2 n(\Delta t)_0}{18\mu K^4} \qquad (n, K \gg 1 \gg |\mathrm{Re}|) \tag{7.49}$$

i.e., $\bar{R}_1(b,n,(\Delta t)_0,K)$ is approximately proportional to $n(\Delta t)_0$ (i.e., the total running time), and $(\Delta x)^4$ if $n, K \gg 1 \gg |\mathrm{Re}|$.

b. Let $\tau^2 \ll 1$. Then with the aid of Eqs. (2.17), (2.18), and (7.2), Eq. (7.30) may be further simplified as

$$R_M(j,n) \doteq \sqrt{A_1^2 + A_2^2}\sin(\phi_j^n + \varepsilon_0) + nO[(2\pi/K)^5] \qquad (\tau^2 \ll 1) \tag{7.50}$$

where

$$A_1 \stackrel{def}{=} \frac{4\pi^3(a-b)(n\Delta t)}{3K^2} \quad , \quad A_2 \stackrel{def}{=} \frac{4\pi^4\mu(n\Delta t)}{3K^2} \tag{7.51}$$

and ε_0 is a number such that

$$\sin\varepsilon_0 = \frac{A_1}{\sqrt{A_1^2+A_2^2}} \quad , \quad \cos\varepsilon_0 = \frac{A_2}{\sqrt{A_1^2+A_2^2}} \tag{7.52}$$

Assuming that the term $n\,O[(2\pi/K)^5]$ in Eq. (7.50) is negligible, then Eqs. (7.45), (7.47), (7.50), and (7.51) imply that

$$\bar{R}_M \doteq \frac{8\pi^2\sqrt{(a-b)^2+\pi^2\mu^2}\,(n\Delta t)}{3K^2} \qquad (K \gg 1 \gg \tau^2) \tag{7.53}$$

i.e., $\bar{R}_M$ is proportional to $n\Delta t$ and $(\Delta x)^2$ if $K \gg 1 \gg \tau^2$.

c. Let $R_1(j,n)$ given in Eq. (7.15) be dominated by the leading term in its principal part. Then it may be shown that

$$\bar{R}_1 \doteq \frac{2n\pi^2\,|\tau\Delta_1(\tau,\delta)|}{3K^3} \qquad (K \gg 1) \tag{7.54}$$

Similarly, let $R_M(j,n)$ given in Eq. (7.30) be dominated by its leading term. Then it may be shown that

$$\bar{R}_M \doteq \frac{8n\pi^2\,|\tau(1-\tau^2)|}{3K^3} \qquad (K \gg 1) \tag{7.55}$$

Let $n\tau \neq 0$. Then the right sides of Eqs. (7.54) and (7.55) will be equal if and only if either *(i)* $(1-\tau^2)^2+\delta^2=0$ or *(ii)* $5(1-\tau^2)^2-3\delta^2=0$. Since $1-\tau^2>0$ and $\delta \geq 0$, case *(i)* cannot occur and case *(ii)* is reduced to

$$1-\tau^2 = \sqrt{3/5}\,\delta \tag{7.56}$$

i.e., $\bar{R}_1$ *and* $\bar{R}_M$ *are approximately equal if* τ *and* δ *are related by Eq. (7.56).* Note that the only difference between Eqs. (5.70) and (7.56) is that the factor $\sqrt{3}$ in the former is replaced by $\sqrt{3/5}$ in the latter.

Let $(a-b)$, μ, and Δx be given. Then it may be shown that there is one and only one set of values of τ, δ, and Δt (denoted by $\bar{\bar{\tau}}_0$, $\bar{\bar{\delta}}_0$, and $(\bar{\bar{\Delta t}})_0$, respectively) such that Eq. (7.56) is satisfied. These values may be determined by using a procedure similar to that used in the determination of τ_0, δ_0, and $(\Delta t)_0$ (see Section 5).

Let μ, Δt, and Δx be given. Eq. (7.56) has no solution if $\delta > \sqrt{5/3}$. If $\sqrt{5/3} \geq \delta > 0$, then Eq. (7.56) is satisfied if

$$\tau = \bar{\bar{\tau}}_{\pm} \stackrel{def}{=} \pm\sqrt{1-\sqrt{3/5}\,\delta} \tag{7.57}$$

d. Let $\tau = 0$. Then the approximations given in Eqs. (7.31), (7.33), and (7.34) may be applicable. In each of them, we further assume that it is dominated by the leading term in its principal part. Then it may be shown that

$$\bar{R}_1 \doteq \frac{n\pi^3\delta\,|1-3\delta^2|}{6K^4} = \frac{2\pi^3\mu(n\Delta t)\,|1-3\delta^2|}{3K^2} \qquad (\tau=0,\ K\gg 1) \tag{7.58}$$

$$\bar{R}_L \doteq \frac{2n\pi^3\delta\,|1-\frac{3}{4}\delta^2|}{3K^4} = \frac{8\pi^3\mu(n\Delta t)\,|1-\frac{3}{4}\delta^2|}{3K^2} \qquad (\tau=0,\ K\gg 1) \tag{7.59}$$

$$\bar{R}_M \doteq \frac{2n\pi^3\delta}{3K^4} = \frac{8\pi^3\mu(n\Delta t)}{3K^2} \qquad (\tau=0,\ K\gg 1) \tag{7.60}$$

Thus $\bar{R}_M$, approximately, is proportional to $n\Delta t$ and $(\Delta x)^2$ if $\tau = 0$. However, one must assume both $\tau = 0$ and $|\delta| \ll 1$ in order that $\bar{R}_1$ and $\bar{R}_L$, be approximately proportional to $n\Delta t$ and $(\Delta x)^2$.

Assuming $\delta > 0$, one may conclude that

(i) $\bar{R}_1 \doteq \bar{R}_L$ if $\delta = \sqrt{5/6} \doteq 0.913$ $(\tau=0,\ K\gg 1)$ (7.61)

(ii) $\bar{R}_1 \doteq \bar{R}_M$ if $\delta = \sqrt{5/3} \doteq 1.291$ $(\tau=0,\ K\gg 1)$ (7.62)

and

(iii) $\bar{R}_L \doteq \bar{R}_M$ if $\delta = \sqrt{8/3} \doteq 1.633$ $(\tau=0,\ K\gg 1)$ (7.63)

This completes the preliminaries for the following numerical evaluation. Several sets of test problems will be considered. In set #1, the test problems have the same values of a (=1), μ (=0.1), b (=0), K (=30), and t (=0.5). They are distinguished by their values of Δt. For each member of set #1, $n = \mathrm{IN}(t/\Delta t)$ where

$$\mathrm{IN}(x) \stackrel{def}{=} \text{the integer nearest to the real number } x. \tag{7.64}$$

As a result, $n\Delta t \doteq t$ if $t/\Delta t \gg 1$. Sets #2 - #4 are defined similarly. The values of a, μ, b, K, t, Re, τ_0, τ_*, and $\bar{\bar{\tau}}_0$ for sets #1 - #4 are listed in Table 7.1. Among these parameters, τ_* is the only one yet to be defined.

For sets #1-#3, $|a| \approx 2\pi\mu$. Thus the corresponding $u_a(x,t)$ is convection-diffusion comparable. For set #4, $|a| \gg 2\pi\mu$. Thus the corresponding $u_a(x,t)$ is convection dominant.

Because *(i)* $\tau \stackrel{def}{=} (a-b)\Delta t/\Delta x$, and *(ii)* a, b, and Δx are constant among the test problems in any one of sets #1 - #4, each test problem within one set can be identified by its value of τ. In

Figs. 7.1 - 7.4, the values of $\bar{E}(b, \mathrm{IN}(t/\Delta t), \Delta t, K)$ for the current, the L/D-F, and the MacCormack schemes are plotted against the values of τ for representative test problems in sets #1 - #4. The following remarks pertain to the results shown in these figures:

a. For the test problems in sets #1 and #2,

$$\mathrm{Re} = \frac{\tau}{\delta} = \frac{1}{12}$$

Thus, on the $\delta - \tau$ plane, each of these problems is represented by a point on the straight line $\tau = \delta/12$. Since $|\mathrm{Re}| \ll 1$, one may conclude from Fig. 4.1 that the MacCormack scheme will become unstable at a τ with $|\tau| \ll 1$. As a matter of fact, it may be estimated that the above straight line intercepts the stability boundary at $\tau \doteq 0.17$. Note that, as a result of how stability is defined (see Section 4), instability generally occurs at a value of τ slightly greater than what is predicted by the stability map if K and n are finite.

For the test problems in set #3, Re = 1/24. Thus the MacCormack scheme becomes unstable at an even smaller τ. On the other hand, $\mathrm{Re} = 5/6 \simeq 1$ for the test problems in set #4. Thus the MacCormack scheme becomes unstable at a τ very close to 1.

b. The most remarkable result shown in each of Figs. 7.1 - 7.4 is the sharp increase in the accuracy of the current method in the neighborhood of $\tau = \tau_0$. For each of sets #1 - #3, which is characterized by $|\mathrm{Re}| \ll 1$, the peak accuracy is 3 - 4 orders of magnitude higher than what can be achieved by either the L/D-F scheme or the MacCormack scheme. For set #4, which is characterized by $\mathrm{Re} \simeq 1$, the peak accuracy of the current method is about a factor of 30 higher than that of the MacCormack method which occurs just before it becomes unstable.

Moreover, for each of sets #1 - #4, the actual value of τ (denoted by τ_* in Table 7.1) at which the peak accuracy of the current method occurs is extremely close to the theoretical value τ_0. The difference is numerically insignificant. As an example, it can be determined numerically that $\bar{E}_0 = -4.923$ and $\bar{E}_* = -4.924$ for set #1. Here $\bar{E}_0$ and $\bar{E}_*$, respectively, denote the values of $\bar{E}$ of the present method for the test problems with $\tau = \tau_0$ and $\tau = \tau_*$.

c. For sets #1 - #3, $|\mathrm{Re}| \ll 1$. Thus $\bar{E}_0$ may be estimated by using Eqs. (7.42) and (7.49). Using these equations, one obtains $\bar{E}_0 = -4.97$ for set #1, $\bar{E}_0 = -4.67$ for set #2, and $\bar{E}_0 = -6.18$ for set #3. For set #4, $|\mathrm{Re}| \simeq 1$ and thus Eq. (7.49) is not applicable. However, by using Eqs. (7.42) and (7.46), one has $\bar{E}_0 = -3.33$ for set #4. From Figs. 7.1 - 7.4, one can infer that the above estimates of $\bar{E}_0$ agree very well with those of $\bar{E}_*$.

d. According to Eqs. (7.42) and (7.53), parameter $\bar{E}$ obtained using the MacCormack scheme is not dependent on Δt (and thus τ) if $(a - b)$, μ, K, and $n\Delta t$ are held constant and $K \gg 1 \gg$

τ^2. This explains the fact that, for the MacCormack scheme, the values of $\bar{E}$ shown in each of Figs. 7.1 - 7.3 are hardly dependent on τ.

Moreover, the approximation of $\bar{R}_M$ given by Eq. (7.53) generally is very accurate. As an example, for the test problems in sets #1 - #3 with $\tau = 0.01$, estimates of $\bar{E}$ obtained by using Eqs. (7.42) and (7.53) are -1.8146, -1.5135, and -2.4166, respectively. On the other hand, the actual numerical values are -1.8148, -1.5119, and -2.4167, respectively. The above comparison also indicates that the accuracy of the approximation of $\bar{R}_M$ improves with the absolute value of $\bar{E}$, i.e., the accuracy of the numerical calculation. The reason for this fact is explained in statements following Eqs. (5.40) and (D.7).

e. It was argued in Section 5 that the current scheme will generally be more accurate than the MacCormack scheme by a factor of 4 when δ is small and the initial condition is smooth. Since $\tau = \mathrm{Re}\cdot\delta$ and Re is a constant in each of sets #1 - #4, one would expect that, as $\tau \to 0$, the value of $|\bar{E}|$ for the current scheme will be greater than that for the MacCormack scheme by approximately $\log_{10} 4 \doteq 0.602$. This expectation is certainly confirmed by the results shown in Figs. 7.1 - 7.4. The same results also indicate that, as $\tau \to 0$, the value of $\bar{E}$ for the MacCormack scheme coincides with that for the L/D-F scheme. This fact may be explained by the following observations: The first term in the principal part of $R_L(j,n)$ given in Eq. (7.26) becomes negligible compared with the second term as $\tau \to 0$ (and $\delta = \tau/\mathrm{Re} \to 0$). On the other hand, the latter is reduced to the leading term in $R_M(j,n)$ (see Eq. (7.30)) as $\delta \to 0$.

f. The values of $\bar{\bar{\tau}}_0$ for sets #1 - #4 are given in Table 7.1. From Figs. 7.1 - 7.4, it is seen that, for each of sets #1 - #4, the value of $\bar{E}$ for the current scheme is approximately equal to that for the MacCormack scheme at $\tau = \bar{\bar{\tau}}_0$. This observation confirms a prediction made in a previous theoretical discussion.

This completes the discussion of the test problems in sets #1 - #4. Next we will study the test problems in sets #5 and #6. As shown in Table 7.2, the test problems in each of these sets share the same values of a, μ, K, n, and Δt. They are different in their values of b. As a result, each member in set #5 or #6 may be identified by its value of τ. Note that, for each test problem in sets #5 and #6, the total running time $n\Delta t = 26/(3\sqrt{3}) \doteq 5.0037$.

Since $a - b = 1.0$ for sets #1 - #4, no member in these sets may have $\tau = 0$. Contrarily, $(a - b)$ is a variable in sets #5 and #6 and therefore both contain a member with $\tau = 0$. Recall that the leading terms in the principal and spurious parts of $R_1(j,n)$, $R_2(j,n)$, $R_L(j,n)$ and $R_M(j,n)$ will be annihilated if $\tau = 0$. In other words, for either of sets #5 and #6, and for all three schemes considered, the test problem with $\tau = 0$ is expected to have the highest accuracy.

The leading terms in the principal parts of $R_1(j,n)$ and $R_2(j,n)$ can also be annihilated if $\tau = \tau_+$ or $\tau = \tau_-$. The parameters τ_+ and τ_- are defined in Eq. (5.77) and the values of τ_+ for sets #5 and #6 are given in Table 7.2. Note that, for set #6, $\tau_+ = \tau_- = 0$. Thus the accuracy peaks of the current scheme that occur at $\tau = 0$, $\tau = \tau_+$, and $\tau = \tau_-$ will merge into one. Furthermore, when $\tau = \tau_+ = \tau_- = 0$, Eq. (7.15) is reduced to

$$R_1(j,n) \doteq n\,O[(2\pi/K)^5] - \frac{1}{576}(2\pi/K)^4 \sin(\phi_j^n) + O[(2\pi/K)^5] \qquad (7.65)$$

i.e., all the explicitly-given leading terms in the principal part vanish. As a result, the approximation of $\bar{R}_1$ given by Eq. (7.58) also vanishes. Obviously, in the case where $\tau_+ = \tau_- = 0$, the estimation of $\bar{E}$ at $\tau = 0$ for the present scheme requires an approximation containing more explicitly-given terms.

In Figs. 7.5 and 7.6, the values of $\bar{E}$ for the current, the L/D-F, and the MacCormack schemes are plotted against the values of τ for representative test problems in set #5 and #6. The following remarks are for the results shown in these figures:

a. On the $\delta - \tau$ plane, each test problem in set #5 is represented by a point on the vertical line $\delta = 0.8/\sqrt{3}$. From Fig. 4.1, one concludes that the MacCormack scheme will be stable for a test problem as long as $|\tau| \lesssim 0.9$. A similar conclusion may be applied to the test problems in set #6. These observations are consistent with the numerical results shown in Figs. 7.5 and 7.6.

b. As expected, the accuracy of all three schemes considered reaches its highest level at $\tau = 0$.

   The values of $\bar{E}$ at $\tau = 0$ can be estimated by using Eqs. (7.42) and (7.58) - (7.60). For the test problem with $\tau = 0$ in set #5, we obtain $\bar{E} \doteq -3.38$, -2.41, and -2.34, respectively, for the current, the L/D-F, and the MacCormack schemes. These estimates agree very well with the results shown in Fig. 7.5.

   For the test problem with $\tau = 0$ in set #6, Eqs. (7.42), (7.59), and (7.60) imply that $\bar{E} \doteq -2.46$ and -2.34, respectively, for the L/D-F and the MacCormack schemes. These estimates are also in good agreement with the results shown in Fig. 7.6. As noted previously, for the test problem considered here, the approximation given in this paper is inadequate in providing an estimate of $\bar{E}$ for the current scheme.

c. In Fig. 7.5, another accuracy peak for the current scheme also appears in the neighborhood of $\tau = \tau_+$. The actual value of τ (denoted by τ_* in Table 7.2) at which the peak accuracy occurs again is very close to the theoretical value τ_+. With the aid of Eqs. (5.77), (7.46), and (7.42), it is estimated that $\bar{E} \doteq -3.04$ for the present scheme at $\tau = \tau_+$. This is very close to the actual peak value one observes in Fig. 7.5. Note that the accuracy peaks at $\tau =$

τ_+ and $\tau = 0$ are merged in Fig. 7.6.

d. From Figs. 7.5 and 7.6, and Table 7.2, one may confirm that, for each of sets #5 and #6, the values of $\bar{E}$ for the current and the MacCormack schemes are approximately equal at $\tau = \bar{\bar{\tau}}_+$.

Next we study the test problems in set #7. According to Table 7.3, these problems share the same values of a, μ, b, K, and t. They differ in their values of Δt. Again we assume that $n = \text{IN}(t/\Delta t)$. Because the relations $\tau = 0$ and $\delta = 36\Delta t$ hold for these problems, the error measure $\bar{E}$ is plotted against δ in Fig. 7.7 for all three schemes considered. From this figure, one observes that

a. The value of $\bar{E}$ for the MacCormack scheme is hardly dependent on δ before it becomes unstable near $\delta \doteq 2.1$.

b. As $\delta \to 0$, *(i)* the value of $\bar{E}$ for the L/D-F scheme approaches that for the MacCormack scheme, and *(ii)* the difference between the value of $\bar{E}$ for the current scheme and that for the other two schemes approaches $\log_{10}4 \doteq 0.602$.

c. The accuracy of the current and the L/D-F schemes has a sharp rise, respectively, near $\delta = 1/\sqrt{3} \doteq 0.577$ and $\delta = 2/\sqrt{3} \doteq 1.155$.

The above observations can be explained by using the stability map Fig. 4.1 and Eqs. (7.58) - (7.60). In addition, the results shown in Fig. 7.7 also confirm the predictions given by Eqs. (7.61) - (7.63).

The last test problems to be considered are those in set #8 (see Table 7.3). These problems share the same values of a, μ, b, K, and t, and differ in their values of Δt. For them, we have $\delta = 0$, $\tau = 30\Delta t$ and $n = \text{IN}(t/\Delta t) = \text{IN}(15/\tau)$. In Fig. 7.8, the values of $\bar{R}_M/\bar{R}_1$ and $\bar{R}_L/\bar{R}_1$ are plotted against τ. One observes that:

a. $\bar{R}_M/\bar{R}_1$ is nearly a constant ($\doteq 4$) throughout the range $1 > \tau > 0$.

b. $\bar{R}_L/\bar{R}_1$ is nearly a constant ($\doteq 4$) in the range $0.4 > \tau > 0$. Its dependence on τ becomes more and more irregular as τ increases from 0.4 to 1.0.

With the aid of Eqs. (7.14) and (7.29), observation (a) can be explained by using Eqs. (5.84) and (5.87). On the other hand, with the aid of Eqs. (7.14) and (7.25), observation (b) can be explained by using Eqs. (5.84) - (5.86) and the fact that $n = \text{IN}(15/\tau)$ is relatively small when $\tau \geq 0.4$.

So far in this section, no numerical results are provided for the error $R_2(j,n)$. This omission is because our discussions have focused on the comparisons of the three schemes considered, and there are no simple counterparts of $R_2(j,n)$ in the MacCormack and the L/D-F schemes. However, it should be noted that the errors $R_1(j,n)$, $R_2(j,n)$, $R_L(j,n)$, and $R_M(j,n)$ at different

(j,n), respectively, have been thoroughly compared against their approximations obtained by evaluating all the explicitly-given terms on the right sides of Eqs. (7.15), (7.22), (7.26), and (7.30). Without going into details, it is sufficient to state that these approximations are highly accurate as long as the errors they approximate are sufficiently small.

8. FINAL REMARKS

Using Eq. (2.2) as a model equation, the basic concepts and properties of the current numerical framework were described and studied in sections 1 - 7. By employing nontraditional discrete variables and taking advantage of the flexibility gained in a unified treatment of space and time, we were able to construct an explicit marching procedure from a single flux conservation principle.

Several fundamental differences that separate the current scheme from other explicit schemes, and how these differences result in greater stability and accuracy for the current scheme were discussed and explained near the end of Section 2. Other important concepts and ideas were discussed in the last part of Section 1.

Perhaps the most intriguing results presented in the current paper are the similarities between the L/D-F scheme and the current scheme. It has been shown that:

a. There is a remarkable similarity between the forms of the amplification factors of these two schemes.

b. These two schemes have the same stability region on the $\delta-\tau$ plane.

c. The stability condition of the current scheme, as in the case of the L/D-F scheme, is essentially the CFL condition and thus independent of the viscosity coefficient μ.

d. Both schemes have no numerical diffusion in the absence of viscosity.

e. The consistency of the current scheme, as in the case of the L/D-F scheme, requires that $\Delta t/\Delta x \rightarrow 0$ as $\Delta t, \Delta x \rightarrow 0$.

Since most of the above similarities are nontrivial, their existence suggests that there may be deeper reasons behind these similarities. Progresses made in a study along this direction will be reported in near future.

Despite of the above similarities, it was shown theoretically and numerically that the current scheme is far superior to the L/D-F scheme in accuracy.

In order to clarify the discussion on consistency, the concept of "strong consistency" is also introduced. It is shown that a scheme which is strongly consistent with the PDE being solved has no intrinsic advantage over a scheme which is not.

This paper is concluded with the following odds and ends:

a. The triangle ΔPQR depicted in Fig. 8.1 is a boundary conservation element. Let α, β, and γ be constants. Let (x_0,t_0) be the coordinates of the mid-point M of $\overline{QR}$. Let Δ''PQR be the interior of ΔPQR. Let

$$\underline{u}(x,t) = \alpha(x-x_0) + \beta(t-t_0) + \gamma \qquad (x,t) \in \Delta''PQR \tag{8.1}$$

be the approximation of $u(x,t)$ in Δ''PQR. Let the flux entering ΔPQR through the edge $\overline{PQ}$ and the value of u at M (denoted by u_M) be given. Then α, β, and γ can be determined by the requirements that *(i)* the total flux leaving ΔPQR = 0, *(ii)* the flux be balanced at $\overline{PQ}$, and *(iii)* $\underline{u}(x_0,t_0) = \gamma = u_M$. Upon determining α, β, and γ, one can calculate the flux leaving ΔPQR through $\overline{RP}$.

b. In Fig. 8.2, a space-time E_2 is divided into conservation elements that are hexagons. Each hexagon has three incoming fluxes and three outgoing fluxes. A marching procedure can be defined if the outgoing fluxes are expressed in terms of the incoming fluxes. A possible way to do this will be described as follows. Let

$$\begin{aligned}\underline{u}(x,t) \stackrel{def}{=}\ & A(x-x_0)^2 + B(x-x_0)(t-t_0) \\ & + C(t-t_0)^2 + D(x-x_0) + E(t-t_0) + F\end{aligned} \tag{8.2}$$

be the approximation of $u(x,t)$ in the interior of a hexagon. Here A, B, C, D, E, and F are constants, and (x_0,t_0) are the coordinates of its center. Let $\vec{\underline{h}}(x,t)$ be defined by Eq. (2.4) and

$$\vec{\nabla} \cdot \vec{\underline{h}}(x,t) = 0 \tag{8.3}$$

Then the divergence theorem implies that the total flux leaving this hexagon = 0. Eqs. (8.2) and (8.3) also imply that

$$B = -2aA \quad , \quad C = a^2 A \quad , \quad E = -aD + 2\mu A \tag{8.4}$$

As a result,

$$\begin{aligned}\underline{u}(x,t) = & A[(x-x_0)-a(t-t_0)]^2 \\ & + D[(x-x_0)-a(t-t_0)] + 2\mu A(t-t_0) + F\end{aligned} \tag{8.5}$$

The coefficients A, D, and F in Eq. (8.5) can be determined in terms of the three incoming fluxes. In turn, the outgoing fluxes can be determined by using Eqs. (2.4) and (8.5).

c. As depicted in Fig. 8.3, a space-time can also be divided into conservation elements of different geometry shapes.

d. A possible conservation element in a space-time E_3 is shown in Fig. 8.4. This conservation element is formed by three "incoming" surfaces RSWV, QRVU, and TWVU, and three "outgoing" surfaces PSWT, PTUQ, and PSRQ. The "marching" direction is that of $\vec{VP}$. Because four points generally are not on a plane in E_3, a surface like RSWV can be formed by two triangles, ΔRWV and ΔRSW.

In the interior of this conservation element, we may assume that

$$\underline{u}(x,y,t) = \alpha(x - x_0) + \beta(y - y_0) + \gamma(t - t_0) + \delta \tag{8.6}$$

where α, β, γ, and δ are constants, and (x_0,y_0,t_0) are the coordinates of the center of the conservation element. The parameters α, β, γ, and δ can be determined by the requirements that *(i)* the total flux leaving the conservation element = 0 and *(ii)* the fluxes be balanced at three incoming surfaces.

Appendix A

In this appendix, stability will be discussed assuming $1-\tau^2+\delta \neq 0$ and $\delta \geq 0$.

Eq. (4.14) implies that

$$\sigma_+(\theta)\cdot\sigma_-(\theta) = \frac{\delta-(1-\tau^2)}{\delta+(1-\tau^2)} \tag{A.1}$$

Thus

$$|\sigma_+(\theta)|\cdot|\sigma_-(\theta)| \begin{cases} >1 & \text{if } \tau^2>1 \text{ and } \delta>0 \\ <1 & \text{if } \tau^2<1 \text{ and } \delta>0 \end{cases} \tag{A.2}$$

An immediate result of Eq. (A.2) is

$$\max\{\,|\sigma_+(\theta)|\,,\,|\sigma_-(\theta)|\,\} > 1 \qquad \text{if } \tau^2>1 \text{ and } \delta>0 \tag{A.3}$$

From Eqs. (4.33), (4.34), and (A.3), one concludes that the current scheme is unstable if $\tau^2>1$ and $\delta>0$. This fact coupled with a result established in subsection 4.2, i.e., the scheme is unstable if $\tau^2>1$ and $\delta=0$, leads to the conclusion that *the current scheme is unstable if* $\tau^2>1$ *and* $\delta\geq 0$.

From Eqs. (4.13) and (4.14), one has

$$\sigma_\pm(\theta) = \cos(\frac{\theta}{2}) \pm i\,|\sin(\frac{\theta}{2})| \qquad \text{if } \tau^2=1 \text{ and } \delta>0 \tag{A.4}$$

Thus

$$|\sigma_+(\theta)| = |\sigma_-(\theta)| = 1 \qquad \text{if } \tau^2=1 \text{ and } \delta>0 \tag{A.5}$$

Because *(i)* $\sigma_+(\theta)=\sigma_-(\theta)$ is a necessary condition for the defectiveness of $Q(\theta)$, *(ii)* in the case where $\tau^2=1$ and $\delta>0$, $\sigma_+(\theta)=\sigma_-(\theta)$ occurs only if $\theta=0$, and *(iii)* the matrix $Q(0)$ is the identity matrix when $\tau^2=1$ and $\delta>0$, the matrices $Q(\theta)$ are nondefective for $\pi\geq\theta>-\pi$ if $\tau^2=1$ and $\delta>0$. As a result, Eqs. (4.33) and (A.5) imply that *the current scheme is stable if* $\tau^2=1$ *and* $\delta>0$. Note that the assumption $1-\tau^2+\delta\neq 0$ excludes the case in which $\tau^2=1$ and $\delta=0$.

It was shown in subsections 4.2 and 4.3 that the current scheme is stable if either *(i)* $\tau^2<1$ and $\delta=0$, or *(ii)* $\tau=0$ and $\delta\geq 0$. Thus the only stability problem left to be discussed is that in which $1>\tau^2>0$ and $\delta>0$.

To proceed, note that Eq. (4.13) implies that

$$[\eta(\theta)]^2 + (1-\tau^2)^2 - \delta^2 = X(\theta) + i\,Y(\theta) \tag{A.6}$$

where $X(\theta)$ and $Y(\theta)$ are *real* functions defined by

$$X(\theta) \stackrel{def}{=} (1-\tau^2)^2 \left[1 - \tau^2 \sin^2(\frac{\theta}{2}) \right] - \delta^2 \sin^2(\frac{\theta}{2}) \tag{A.7}$$

$$Y(\theta) \stackrel{def}{=} -2\delta\tau(1-\tau^2) \sin(\frac{\theta}{2}) \cos(\frac{\theta}{2}) \tag{A.8}$$

In the following derivations, $X(\theta)$ and $Y(\theta)$, respectively, may be abbreviated as X and Y. Since, by definition, the range of the phase angle of the principal square root is $(-\pi/2\ ,\ \pi/2]$, we have

$$\sqrt{X + iY} = \frac{1}{\sqrt{2}} \left[\sqrt{\sqrt{X^2+Y^2} + X} + i\, \mathrm{sign}(Y) \sqrt{\sqrt{X^2+Y^2} - X} \right] \tag{A.9}$$

where $\mathrm{sign}(Y) \stackrel{def}{=} 1$ if $Y \geq 0$ and $\mathrm{sign}(Y) \stackrel{def}{=} -1$ if $Y < 0$. By using Eqs. (4.14), (A.6) and (A.9), one concludes that

$$\sigma_{\pm}(\theta) = \left\{ \left[\delta \cos(\frac{\theta}{2}) \pm \frac{1}{\sqrt{2}} \sqrt{\sqrt{X^2+Y^2} + X} \right] \right.$$

$$\left. + i \left[\pm \frac{1}{\sqrt{2}} \mathrm{sign}(Y) \sqrt{\sqrt{X^2+Y^2} - X} - \tau(1-\tau^2) \sin(\frac{\theta}{2}) \right] \right\} \Big/ (1 - \tau^2 + \delta) \tag{A.10}$$

It will now be shown that

$$|\sigma_+(\theta)| < 1 \quad \text{and} \quad |\sigma_-(\theta)| < 1 \qquad \text{if } \tau^2 < 1\ ,\ \delta > 0 \text{ and } \theta \neq 0 \tag{A.11}$$

Proof: Using Eq. (A.10), one has

$$\left[|\sigma_+(\theta)|^2 + |\sigma_-(\theta)|^2 \right] (1 - \tau^2 + \delta)^2$$

$$= 2 \left[\delta^2 \cos^2(\frac{\theta}{2}) + \tau^2 (1-\tau^2)^2 \sin^2(\frac{\theta}{2}) + \sqrt{X^2+Y^2} \right] \tag{A.12}$$

Combining Eqs. (A.1) and (A.12), one obtains

$$\left[1 - |\sigma_+(\theta)|^2 \right] \left[1 - |\sigma_-(\theta)|^2 \right] (1 - \tau^2 + \delta)^2$$

$$= 2 \left\{ (1-\tau^2)^2 \left[1 - \tau^2 \sin^2(\frac{\theta}{2}) \right] + \delta^2 \sin^2(\frac{\theta}{2}) - \sqrt{X^2+Y^2} \right\} \tag{A.13}$$

Because *(i)*

$$\left\{ (1-\tau^2)^2 \left[1 - \tau^2 \sin^2(\frac{\theta}{2}) \right] + \delta^2 \sin^2(\frac{\theta}{2}) - \sqrt{X^2+Y^2} \right\}$$

$$\times \{ (1-\tau^2)^2 [1-\tau^2 \sin^2(\frac{\theta}{2})] + \delta^2 \sin^2(\frac{\theta}{2}) + \sqrt{X^2+Y^2} \}$$

$$\equiv 4\delta^2 (1-\tau^2)^3 \sin^2(\frac{\theta}{2}) \qquad \text{(A.14)}$$

(ii)

$$(1-\tau^2)^2 [1-\tau^2 \sin^2(\frac{\theta}{2})] + \delta^2 \sin^2(\frac{\theta}{2}) + \sqrt{X^2+Y^2} > 0 \qquad \text{if } \tau^2 < 1 \qquad \text{(A.15)}$$

and *(iii)*

$$4\delta^2 (1-\tau^2)^3 \sin^2(\frac{\theta}{2}) > 0 \qquad \text{if } \tau^2 < 1 \ , \ \delta > 0 \text{ and } \theta \neq 0 \ , \qquad \text{(A.16)}$$

one concludes that the expression on the the right side of Eq. (A.13) > 0 if $\tau^2 < 1$, $\delta > 0$ and $\theta \neq 0$. Thus

$$[1 - |\sigma_+(\theta)|^2] [1 - |\sigma_-(\theta)|^2] > 0 \qquad \text{if } \tau^2 < 1 \ , \ \delta > 0 \text{ and } \theta \neq 0 \qquad \text{(A.17)}$$

According to Eq. (A.2), $|\sigma_+(\theta)| \cdot |\sigma_-(\theta)| < 1$ if $\tau^2 < 1$ and $\delta > 0$. This inequality combined with Eq. (A.17) implies Eq. (A.11). Q.E.D.

Note that *(i)*

$$|\sigma_+(0)| = 1 \quad \text{and} \quad |\sigma_-(0)| = \left| \frac{\delta - (1-\tau^2)}{\delta + (1-\tau^2)} \right| < 1 \qquad \text{if } \tau^2 < 1 \text{ and } \delta > 0 \qquad \text{(A.18)}$$

and *(ii)* $Q(0)$ is diagonal and thus nondefective. According to Eqs. (4.33) and (4.34), (i), (ii), and Eq. (A.11) imply that the current scheme is stable if $\tau^2 < 1$ and $\delta > 0$. Combining this and other results obtained earlier, one concludes that, assuming $\delta \geq 0$ and $1 - \tau^2 + \delta \neq 0$, the current scheme is stable if and only if $\tau^2 \leq 1$.

Appendix B

The main L/D-F scheme is defined in Eq. (3.19). It will be supplemented by the starting scheme

$$\frac{u_j^1 - u_j^0}{\Delta t} + (a-b)\frac{u_{j+1}^0 - u_{j-1}^0}{2\Delta x} - \mu \frac{u_{j+1}^0 + u_{j-1}^0 - 2u_j^0}{(\Delta x)^2} = 0 \tag{B.1}$$

Because the moving mesh shown in Fig. 2.1(a) is used in the current discussion, coefficient a is replaced by $a-b$ in Eq. (B.1). We will also assume periodic conditions:

$$u_j^n = u_{j+K}^n \qquad (j = 0, \pm 1, \pm 2, \cdots, n = 0, 1, 2, \cdots) \tag{B.2}$$

As a result of Eq. (B.2),

$$u_j^n = \sum_{k=0}^{K-1} \underline{u}_k^n \, \phi_j^{(k)} \tag{B.3}$$

where $\phi_j^{(k)}$ are defined by Eq. (4.2) and

$$\underline{u}_k^n \stackrel{def}{=} \sum_{l=0}^{K-1} u_l^n \, \overline{\phi_l^{(k)}} \tag{B.4}$$

Substituting Eq. (B.3) into Eq. (3.19) and using Eqs. (2.17), (2.18), and (4.7), one obtains

$$(1 + \frac{\delta}{2})\,\underline{u}_k^{n+1} + (2i\tau\sin\theta_k - \delta\cos\theta_k)\,\underline{u}_k^n - (1 - \frac{\delta}{2})\,\underline{u}_k^{n-1} = 0 \quad , \quad n = 1, 2, 3, \cdots \tag{B.5}$$

Note that Eq. (B.5) can be expressed as

$$\vec{\underline{U}}_k^{n+1} = M(\theta_k)\,\vec{\underline{U}}_k^n \quad , \quad n = 0, 1, 2, \cdots \tag{B.6}$$

where

$$\vec{\underline{U}}_k^n \stackrel{def}{=} \begin{bmatrix} \underline{u}_k^{n+1} \\ \\ \underline{u}_k^n \end{bmatrix} \tag{B.7}$$

and

$$M(\theta) = \begin{bmatrix} \dfrac{\delta\cos\theta - 2i\tau\sin\theta}{1 + (\delta/2)} & \dfrac{1 - (\delta/2)}{1 + (\delta/2)} \\ \\ 1 & 0 \end{bmatrix} \tag{B.8}$$

As a result of Eq. (B.6), $M(\theta_k)$, $k = 0, 1, 2, \cdots, K-1$, can be referred to as the amplification matrices of the L/D-F scheme. Also we have

$$\vec{\underline{U}}_k^n = [M(\theta_k)]^n \, \vec{\underline{U}}_k^0 \quad , \quad n = 0, 1, 2, \cdots \tag{B.9}$$

The eigenvalues of $M(\theta)$, i.e.,

$$A_{\pm}(\theta) = \frac{(\frac{\delta}{2})\cos\theta - i\tau\sin\theta \pm \sqrt{(\frac{\delta}{2}\cos\theta - i\tau\sin\theta)^2 + 1 - (\frac{\delta}{2})^2}}{1 + \frac{\delta}{2}} \tag{B.10}$$

will be referred to as the amplification factors of the L/D-F scheme.

Because $M(\theta)$ is *not* diagonal, it is defective if and only if $A_+(\theta) = A_-(\theta)$. With the aid of this fact and Jordan's theorem [p.362, 6], Eqs. (B.7) and (B.9) can be used to show that

$$\underline{u}_k^n = \begin{cases} h_{k+}[A_+(\theta_k)]^n + h_{k-}[A_-(\theta_k)]^n & \text{if } A_+(\theta_k) \neq A_-(\theta_k) \\ h_{k+}[A_+(\theta_k)]^n + n\,h_{k-}[A_-(\theta_k)]^{n-1} & \text{if } A_+(\theta_k) = A_-(\theta_k) \end{cases} \tag{B.11}$$

where h_{k+} and h_{k-} are constants to be determined by $\vec{U}_k^0$. In this paper, the L/D-F scheme is said to be stable if and only if, for any $K \geq 3$ and any specification of h_{k+} and h_{k-}, $k = 0, 1, 2, \cdots, K-1$ (i.e., any specification of u_j^0 and u_j^1, $j = 0, 1, 2, \cdots, K-1$), u_j^n, $j = 0, \pm1, \pm2, \cdots$, remain bounded as $n \to +\infty$ with the parameters τ and δ being held constant. From Eq. (B.11), one can conclude that the L/D-F scheme is stable if and only if, for any $K \geq 3$ and any $k = 0, 1, 2, \cdots, K-1$, we have

$$\max\{\,|A_+(\theta_k)|\,,\,|A_-(\theta_k)|\,\} \leq 1 \qquad \text{if } A_+(\theta_k) \neq A_-(\theta_k) \tag{B.12}$$

and

$$|A_+(\theta_k)| < 1 \qquad \text{if } A_+(\theta_k) = A_-(\theta_k) \tag{B.13}$$

In the following derivations, we shall prove this: In the case where $\delta > 0$, the L/D-F scheme is stable if and only if $\tau^2 \leq 1$. In the case where $\delta = 0$, it is stable if and only if $\tau^2 < 1$.

Proof: It is easy to show that

$$\max\{\,|A_+(\tfrac{\pi}{2})|\,,\,|A_-(\tfrac{\pi}{2})|\,\} = \frac{|\tau| + \sqrt{\tau^2 - 1 + (\delta/2)^2}}{1 + (\delta/2)} > 1 \quad \text{if } \tau^2 > 1 \text{ and } \delta \geq 0 \tag{B.14}$$

and

$$A_+(\tfrac{\pi}{2}) = A_-(\tfrac{\pi}{2}) \quad \text{and} \quad |A_+(\tfrac{\pi}{2})| = |A_-(\tfrac{\pi}{2})| = 1 \qquad \text{if } \tau^2 = 1 \text{ and } \delta = 0 \tag{B.15}$$

Because $\theta_k = \pi/2$ if $K = 4, 8, 12, \cdots$, and $k = K/4$, Eqs. (B.12) - (B.15) imply that the L/D-F scheme is unstable if either *(i)* $\tau^2 > 1$ and $\delta > 0$, or *(ii)* $\tau^2 \geq 1$ and $\delta = 0$.

Next we observe that

$$\left.\begin{array}{l} |A_+(\theta)| \;=\; |A_-(\theta)| \;=\; 1 \\ \text{and} \\ \qquad\qquad A_+(\theta) \neq A_-(\theta) \end{array}\right\} \quad \text{if } \tau^2 < 1 \text{ and } \delta = 0 \qquad \text{(B.16)}$$

It follows from Eq. (B.12) that the L/D-F scheme is stable if $\tau^2 < 1$ and $\delta = 0$.

The proof is completed by showing that the L/D-F scheme is stable if $\tau^2 \le 1$ and $\delta > 0$. To proceed, from Eq. (B.10), we obtain

$$A_+(\theta)\, A_-(\theta) \;=\; \frac{\delta/2 - 1}{\delta/2 + 1} \qquad \text{(B.17)}$$

and

$$A_\pm(\theta) = \left\{ \left[\frac{\delta}{2}\cos\theta \pm \frac{1}{\sqrt{2}} \sqrt{\sqrt{X^2 + Y^2} + X} \right] \right.$$
$$\left. + \, i \left[\pm \frac{1}{\sqrt{2}} \,\mathrm{sign}(Y) \sqrt{\sqrt{X^2 + Y^2} - X} \;-\; \tau \sin\theta \right] \right\} \Big/ (1 \,+\, \delta/2) \qquad \text{(B.18)}$$

where X and Y, respectively, are the abbreviations of

$$X(\theta) \overset{def}{=} (1 - \tau^2 \sin^2\theta) - (\frac{\delta}{2})^2 \sin^2\theta \quad , \quad \text{and} \quad Y(\theta) \overset{def}{=} -\delta\tau \sin\theta\cos\theta \qquad \text{(B.19)}$$

Also $\mathrm{sign}(Y) \overset{def}{=} 1$ if $Y \ge 0$ and $\mathrm{sign}(Y) \overset{def}{=} -1$ if $Y < 0$. Eq. (B.18) implies that

$$[\, |A_+(\theta)|^2 + |A_-(\theta)|^2 \,] \,(1 + \frac{\delta}{2})^2 \;=\; 2\,[\,(\frac{\delta}{2})^2 \cos^2\theta + \tau^2 \sin^2\theta + \sqrt{X^2 + Y^2}\,] \qquad \text{(B.20)}$$

By using Eqs. (B.17) and (B.20), one concludes that

$$[\,1 - |A_+(\theta)|^2\,]\,[\,1 - |A_-(\theta)|^2\,]\,(1 + \frac{\delta}{2})^2$$
$$= \; 2\,[\,1 - \tau^2 \sin^2\theta + (\frac{\delta}{2})^2 \sin^2\theta - \sqrt{X^2 + Y^2}\,] \qquad \text{(B.21)}$$

Note that

$$[\,1 - \tau^2 \sin^2\theta + (\frac{\delta}{2})^2 \sin^2\theta - \sqrt{X^2 + Y^2}\,]\,[\,1 - \tau^2 \sin^2\theta + (\frac{\delta}{2})^2 \sin^2\theta + \sqrt{X^2 + Y^2}\,]$$
$$\equiv \; \delta^2 (1 - \tau^2) \sin^2\theta \qquad \text{(B.22)}$$

and

$$[\,1-\tau^2\sin^2\theta+(\frac{\delta}{2})^2\sin^2\theta+\sqrt{X^2+Y^2}\,]\;>\;0 \qquad \text{if } \tau^2\le 1 \text{ and } \delta>0 \qquad \text{(B.23)}$$

Combining Eqs. (B.21) - (B.23), one concludes that

$$[\,1-|A_+(\theta)|^2\,]\,[\,1-|A_-(\theta)|^2\,]=0 \qquad \text{if } \tau^2=1 \text{ and } \delta>0 \qquad \text{(B.24)}$$

and

$$[\,1-|A_+(\theta)|^2\,]\,[\,1-|A_-(\theta)|^2\,]>0 \qquad \text{if } \tau^2<1\,,\ \delta>0 \text{ and } \sin\theta\neq 0 \qquad \text{(B.25)}$$

An immediate result of Eq. (B.17) is

$$|A_+(\theta)|\cdot|A_-(\theta)|<1 \qquad \text{if } \delta>0 \qquad \text{(B.26)}$$

From Eqs. (B.24) - (B.26), one concludes that *(i)*

$$\left.\begin{array}{l}\max\{\,|A_+(\theta)|\,,\,|A_-(\theta)|\,\}=1 \\ \text{and} \\ |A_+(\theta)|\neq|A_-(\theta)|\end{array}\right\} \quad \text{if } \tau^2=1 \text{ and } \delta>0 \qquad \text{(B.27)}$$

and *(ii)*

$$|A_+(\theta)|<1 \quad \text{and} \quad |A_-(\theta)|<1 \qquad \text{if } \tau^2<1\,,\ \delta>0 \text{ and } \sin\theta\neq 0 \qquad \text{(B.28)}$$

Furthermore, Eq. (B.10) implies that

$$\left.\begin{array}{l}\max\{\,|A_+(\theta)|\,,\,|A_-(\theta)|\,\}=1 \\ \text{and} \\ |A_+(\theta)|\neq|A_-(\theta)|\end{array}\right\} \quad \text{if } \sin\theta=0 \text{ and } \delta>0 \qquad \text{(B.29)}$$

By using Eqs. (B.12), (B.13), and (B.27) - (B.29), one reaches the conclusion that the L/D-F scheme is stable if $\tau^2\le 1$ and $\delta>0$. Q.E.D.

Next we shall study the accuracy of the numerical solutions obtained by using the L/D-F scheme. Let

$$h(\theta)\ \stackrel{def}{=}\ \frac{A_+(\theta)\;-\;1\;+\;\frac{\delta}{2}(1-\cos\theta)\;+\;i\,\tau\sin\theta}{A_+(\theta)\;-\;A_-(\theta)} \qquad (A_+(\theta)\neq A_-(\theta)) \qquad \text{(B.30)}$$

and

$$\underline{u}_L(j,n,k)\ \stackrel{def}{=}\ b_k\,e^{ij\theta_k}\,\{\,[\,1-h(\theta_k)\,]\,[A_+(\theta_k)\,]^n\;+\;h(\theta_k)\,[A_-(\theta_k)\,]^n\,\}$$

$$(A_+(\theta_k)\neq A_-(\theta_k)) \qquad \text{(B.31)}$$

where b_k are defined in Eq. (5.8). With the aid of Eqs. (B.3) and (B.11), and the assumption that

$$A_+(\theta_k) \neq A_-(\theta_k) \ , \quad k = 0, 1, 2, ..., K-1 \tag{B.32}$$

it can be shown that the solution to the finite-difference problem defined by Eqs. (3.19), (B.1), and (B.2) is given by

$$u_j^n = \underline{u}_L(j,n) \stackrel{def}{=} \sum_{k=0}^{K-1} \underline{u}_L(j,n,k) \tag{B.33}$$

Note that

$$A_+(0) = 1 \quad \text{and} \quad A_-(0) = -\frac{1-\delta/2}{1+\delta/2} \tag{B.34}$$

Thus there is a neighborhood of $\theta = 0$ on the complex θ-plane in which $A_+(\theta) \neq A_-(\theta)$ and thus $h(\theta)$ is defined.

Combining Eqs. (5.53) and (B.31), one has

$$\underline{u}_L(j,n,k) - u_a(j,n,k) = b_k e^{ij\theta_k} \Delta_L(n,\theta_k) \tag{B.35}$$

where

$$\Delta_L(n,\theta) \stackrel{def}{=} [1-h(\theta)]\{[A_+(\theta)]^n - [A_a(\theta)]^n\} + h(\theta)\{[A_-(\theta)]^n - [A_a(\theta)]^n\} \tag{B.36}$$

In the following, we shall study $\Delta_L(n,\theta)$ in the neighborhood of $\theta = 0$.

Eqs. (B.10) and (B.30) imply that

$$h(\theta) = \frac{1}{4}\left(1-\frac{\delta^2}{4}\right)\left\{-\tau^2\theta^2 + \frac{i\tau\delta}{2}(1-2\tau^2)\theta^3 + \frac{1}{48}[4\tau^2(4-9\tau^2) + 3\delta^2(15\tau^4-12\tau^2+1)]\theta^4 + O(\theta^5)\right\} \tag{B.37}$$

Eq. (B.37) is reduced to

$$h(\theta) = -\frac{\tau^2}{4}\theta^2 + \frac{\tau^2(4-9\tau^2)}{48}\theta^4 + O(\theta^5) \qquad \text{if } \delta = 0 \tag{B.38}$$

and

$$h(\theta) = \frac{1}{64}\delta^2\left(1-\frac{\delta^2}{4}\right)\theta^4 + O(\theta^5) \qquad \text{if } \tau = 0 \tag{B.39}$$

According to Eq. (5.12), the sum of the upper elements in $\vec{H}_+(\theta)$ and $\vec{H}_-(\theta)$ is 1. Thus the role played by the upper element of $\vec{H}_-(\theta)$ in Eq. (5.21) is similar to that played by $h(\theta)$ in Eq. (B.36). An inspection of Eqs. (5.32) - (5.34) and (B.37) - (B.39) reveals that the upper element in $\vec{H}_-(\theta)$

is *(i)* smaller than $h(\theta)$ by one order of θ if $\tau\delta \neq 0$, or *(ii)* smaller than $h(\theta)$ by two orders of θ if $\delta = 0$, or *(iii)* in the same order of θ with $h(\theta)$ if $\tau = 0$. Thus, if $\tau \neq 0$, the influence of the spurious part is less noticeable in the current scheme than in the L/D-F scheme. Note that the form of $h(\theta)$ given in Eq. (B.30) is dependent on the forms of both the main scheme Eq. (3.19) and the starting scheme Eq. (B.1). It is an accident of this combined influence of the main and starting schemes that $h(\theta)$ and the upper element in $\vec{H}_-(\theta)$ are in the same order of θ if $\tau = 0$, even though $\partial u / \partial t$ is approximated by a one-sided difference formula in Eq. (B.1).

Next, with the aid of Eqs. (5.17) and (B.10), it can be shown that

$$\varepsilon_{L+}(\theta) \stackrel{def}{=} \frac{A_+(\theta)}{A_a(\theta)} - 1 = \frac{\tau^2\delta}{4}\theta^2 + \frac{i\tau}{6}(1-\tau^2)(1-\frac{3}{4}\delta^2)\theta^3$$

$$+ \frac{\delta}{96}\left\{(1-\tau^2)[\frac{3\tau^2}{2}(5\delta^2-2\delta+4) + 2(1-\frac{3}{4}\delta^2)] + 3\delta\tau^2\right\}\theta^4 + O(\theta^5) \quad \text{(B.40)}$$

and

$$\varepsilon_{L-}(\theta) \stackrel{def}{=} \frac{A_-(\theta)}{-(\frac{1-\delta/2}{1+\delta/2})A_a(\theta)} - 1 = 2i\tau\theta + [\frac{\delta}{4}(2-\tau^2) - 2\tau^2]\theta^2 + O(\theta^3) \quad \text{(B.41)}$$

Obviously, the roles played by $\varepsilon_{L+}(\theta)$ and $\varepsilon_{L-}(\theta)$ in the L/D-F scheme, respectively, are similar to those played by $\varepsilon_+(\theta)$ and $\varepsilon_-(\theta)$ in the current scheme. A comparison between Eqs. (5.36) and (B.40) reveals that, $[\sigma_+(\theta)]^2$ approximates $A_a(\theta)$ to the second order in θ while $A_+(\theta)$ approximates $A_a(\theta)$ only to the first order in θ. Note that the amplification factors are completely determined by the main marching scheme. Thus, the above difference in accuracy has nothing to do with how the starting scheme is defined.

By using the arguments that were used to establish Eqs. (5.40) and (5.45), it can be shown that *(i)*

$$[A_+(\theta)]^n - [A_a(\theta)]^n \doteq n[A_a(\theta)]^n \varepsilon_{L+}(\theta) \quad \text{(B.42)}$$

if $n|\varepsilon_{L+}(\theta)| \ll 1$, and *(ii)*

$$[A_-(\theta)]^n - [A_a(\theta)]^n \doteq -[A_a(\theta)]^n \quad \text{(B.43)}$$

if $\delta > 0$, $|\theta|$ is sufficiently small, and n is sufficiently large.

Let

$$r_{L+}(n,k) \stackrel{def}{=} \frac{[1-h(\theta)]\{[A_+(\theta)]^n - [A_a(\theta)]^n\}}{[A_a(\theta)]^n} \quad \text{(B.44)}$$

and

$$r_{L-}(n,k) \stackrel{def}{=} \frac{h(\theta)\{[A_-(\theta)]^n - [A_a(\theta)]^n\}}{[A_a(\theta)]^n} \tag{B.45}$$

Then Eqs. (5.61) - (5.63) follow directly from Eqs. (5.53), (5.60), (B.35) - (B.37), (B.40), (B.42), and (B.43).

For the special case in which $\delta = 0$, Eq. (B.43) is not applicable. However, since $|A_-(\theta)| = |A_a(\theta)| = 1$ if $\tau^2 < 1$ and $\delta = 0$, one has

$$\left|\frac{[A_-(\theta)]^n - [A_a(\theta)]^n}{[A_a(\theta)]^n}\right| \leq 2 \qquad \text{if } \tau^2 < 1 \text{ and } \delta = 0 \tag{B.46}$$

With the aid of Eqs. (B.38) and (B.45), Eq. (B.46), in turn, can be used to obtain Eq. (5.85).

Appendix C

Eqs. (5.26) - (5.28) and the assumption $n|\varepsilon_+(\theta)| \ll 1$ of Eq. (5.40) will be discussed in this appendix.

Assuming Eqs. (5.23) and (5.24), first we shall show that

a. $\omega(\theta) \neq 0$ if $\pi > |\theta|$; and

b. $\omega(\pi) = 0$ if and only if $1 - \tau^2 - \delta = 0$ and $1 > \tau \geq 0$

Proof: By using Eqs. (4.13), (4.14), (5.23), and (5.24), it is easy to show that

$$\omega(0) = \frac{2(1-\tau^2)}{1-\tau^2+\delta} > 0 \qquad \text{(C.1)}$$

Let I be the 2×2 identity matrix. Then the determinant of the matrix $[Q(\theta) - \sigma_-(\theta)\,\mathrm{I}]$ vanishes. This fact coupled with Eqs. (4.12), (5.23), (5.24), and (C.1) implies that

$$\omega(\theta) \neq 0 \qquad \text{if} \quad \pi \geq \theta > -\pi \quad \text{and} \quad 1 - \tau^2 - \delta \neq 0 \qquad \text{(C.2)}$$

Let $1-\tau^2-\delta = 0$. Then $\delta = 1-\tau^2 > 0$. This inequality combined with Eq. (4.13) implies that the real part of $\eta(\theta)$ is positive if $\pi > |\theta|$. As a result, the principal square root $\sqrt{[\eta(\theta)]^2} = \eta(\theta)$ if $\pi > |\theta|$. Combining Eqs. (4.14) and (5.24), one can conclude that $\sigma_-(\theta) = 0$, and

$$\omega(\theta) = \cos(\theta/2) - i\tau\sin(\theta/2) \neq 0\,, \quad \text{if } \pi > |\theta| \text{ and } 1-\tau^2-\delta = 0 \qquad \text{(C.3)}$$

Statement (a) ia a result of Eqs. (C.2) and (C.3). Statement (b) follows directly from Eq. (C.2) and the fact that

$$\omega(\pi) = (i/2)(|\tau| - \tau) \qquad \text{if } 1-\tau^2-\delta = 0 \qquad \text{(C.4)}$$

Q.E.D.

Next we shall prove Eqs. (5.26) - (5.28) for any θ with $\pi \geq \theta > -\pi$ and $\sigma_+(\theta) \neq \sigma_-(\theta)$. To proceed, note that $\sigma_+(\theta) + \sigma_-(\theta) =$ the trace of $Q(\theta)$. By using Eqs. (4.12) and (5.24), one obtains

$$\omega(\theta) = \left(\frac{1-\tau^2-\delta}{1-\tau^2+\delta}\right)[\cos(\theta/2) + i\tau\sin(\theta/2)] + \sigma_+(\theta) \qquad (\pi \geq \theta > -\pi) \qquad \text{(C.5)}$$

Because $\sigma_+(\theta) \neq \sigma_-(\theta)$, $\vec{g}_+(\theta)$ and $\vec{g}_-(\theta)$, respectively, are the eigenvectors of $Q(\theta)$ with eigenvalues $\sigma_+(\theta)$ and $\sigma_-(\theta)$. With the aid of Eqs. (4.12), (4.22), (5.24), (5.25), and (C.5), it can be shown that

$$g_{21}(\theta) = i\xi_1(\theta)\, g_{11}(\theta) \qquad \text{(C.6)}$$

and

$$g_{12}(\theta) = i\xi_2(\theta)\, g_{22}(\theta) \tag{C.7}$$

By choosing $g_{11}(\theta) = g_{22}(\theta) = 1$, Eqs. (4.21), (C.6), and (C.7) imply that

$$G(\theta) = \begin{bmatrix} 1 & i\xi_2(\theta) \\ i\xi_1(\theta) & 1 \end{bmatrix} \tag{C.8}$$

Because $\vec{g}_+(\theta)$ and $\vec{g}_-(\theta)$ must be linearly independent (i.e., $G(\theta)$ must be nonsingular) if $\sigma_+(\theta) \neq \sigma_-(\theta)$, Eqs. (5.26) - (5.28) follow from Eqs. (C.8) and (5.11) immediately.

As a preliminary for a discussion on the assumption $n\,|\epsilon_+(\theta)| \ll 1$ of Eq. (5.40), we shall establish the following inequality:

Lemma. Let x be a complex number. Let $n > 0$ be an integer. Then

$$|(1+x)^l - 1| \ll 1\ , \quad l = 1, 2, 3, \ldots, n \tag{C.9}$$

if and only if

$$n\,|x| \ll 1 \tag{C.10}$$

Proof: Let the real numbers r, ϕ, ρ, and ψ be defined by the conditions:

$$x = r e^{i\phi}\ , \quad r \geq 0\ , \quad \pi \geq \phi > -\pi \tag{C.11}$$

and

$$1 + x = \rho e^{i\psi}\ , \quad \rho \geq 0\ , \quad \pi \geq \psi > -\pi \tag{C.12}$$

(See Fig. C.1). It follows from Eq. (C.12) that

$$|(1+x)^l - 1| = \sqrt{(\rho^l - 1)^2 + 2\rho^l\,(1 - \cos(l\psi))} \tag{C.13}$$

By using Eq. (C.13) and the fact that $1 - \cos(l\psi) \geq 0$, one concludes that Eq. (C.9) is true if and only if

$$(\rho^l - 1)^2 \ll 1\ , \quad l = 1, 2, 3, \ldots, n \tag{C.14}$$

and

$$1 - \cos(l\psi) \ll 1\ , \quad l = 1, 2, 3, \ldots, n \tag{C.15}$$

Note that Eq. (C.14) implies that ρ^l is very close to 1 for $l = 1, 2, 3, \ldots, n$. As a result, the second term under the radical sign on the right side of Eq. (C.13) will be small compared with 1 for $l = 1, 2, 3, \ldots, n$ if and only if Eq. (C.15) is true. Because $\pi \geq \psi > -\pi$, Eq. (C.15) is equivalent to

$$n\,|\psi| \ll 1 \tag{C.16}$$

Also, Eq. (C.14) is equivalent to

$$|\varepsilon| \ll 1 \quad \text{where } \varepsilon \stackrel{def}{=} \rho^n - 1 \tag{C.17}$$

A result of Eq. (C.17) is

$$\rho = (1+\varepsilon)^{\frac{1}{n}} \doteq 1 + \frac{\varepsilon}{n} \tag{C.18}$$

With the above preparations, first we shall show that Eq. (C.10) is a result of Eq. (C.9). According to Fig. C.1,

$$|x| = r = \sqrt{\rho^2 + 1 - 2\rho\cos\psi} \tag{C.19}$$

By using Eqs. (C.16) - (C.19), one concludes that

$$n\,|x| \doteq n\sqrt{\rho^2 + 1 - 2\rho(1-\psi^2/2)} = n\sqrt{(\rho-1)^2 + \rho\psi^2} \doteq n\sqrt{(\varepsilon/n)^2 + (1+\varepsilon/n)\psi^2}$$
$$= \sqrt{\varepsilon^2 + (1+\varepsilon/n)(n\psi)^2} \ll 1$$

i.e., Eq. (C.10) is true.

Next, assuming Eq. (C.10), we shall prove Eq. (C.9) by induction. Obviously $|(1+x)^l - 1| = |x| \ll 1$ if $l = 1$. Let l_0 be an integer such that $n > l_0 \geq 1$ and

$$|(1+x)^l - 1| \ll 1\,, \quad l = 1, 2, 3, \ldots, l_0 \tag{C.20}$$

As a result of Eq. (C.20), we have

$$|(1+x)^l| < 2\,, \quad l = 1, 2, 3, \ldots, l_0 \tag{C.21}$$

With the aid of Eqs. (C.10) and (C.21), we have

$$|(1+x)^{l_0+1} - 1| \equiv |x\sum_{l=0}^{l_0}(1+x)^l| \leq |x|\sum_{l=0}^{l_0}|1+x|^l < (2l_0+1)\,|x| \ll 1$$

Q.E.D.

With the aid of the above lemma, we now show that

$$[\sigma_+(\theta)]^{2l} \doteq [A_a(\theta)]^l\,, \quad l = 1, 2, \ldots, n \quad (n > 0) \tag{C.22}$$

if and only if

$$n\,|\varepsilon_+(\theta)| \ll 1 \tag{C.23}$$

Proof: By replacing n with l in Eq. (5.38) and recalling the definition of the sign "$\doteq$" given in Section 5, it is seen that Eq. (C.22) is true if and only if

$$|[1+\varepsilon_{+}(\theta)]^{l}-1| \ll 1 \ , \quad l=1, 2, 3, ..., n \tag{C.24}$$

According to the above lemma, Eq. (C.24) is true if and only if Eq. (C.23) is true. Q.E.D.

Appendix D

A version of the MacCormack scheme for Eq. (2.2) [p.163, 3] is

$$\text{Predictor:} \quad \overline{u_j^{n+1}} = u_j^n - \frac{(a-b)\Delta t}{\Delta x}(u_{j+1}^n - u_j^n) + \frac{\mu \Delta t}{(\Delta x)^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) \tag{D.1}$$

$$\text{Corrector:} \quad u_j^{n+1} = \frac{1}{2}\left[u_j^n + \overline{u_j^{n+1}} - \frac{(a-b)\Delta t}{\Delta x}(\overline{u_j^{n+1}} - \overline{u_{j-1}^{n+1}}) + \frac{\mu \Delta t}{(\Delta x)^2}(\overline{u_{j+1}^{n+1}} - 2\overline{u_j^{n+1}} + \overline{u_{j-1}^{n+1}})\right] \tag{D.2}$$

Because the moving mesh depicted in Fig. 2.1(a) is used in the current discussion, coefficient a is replaced by $a-b$ in Eqs. (D.1) and (D.2). In the above version, a forward difference is employed in the predictor step for $\partial u/\partial x$ and a backward difference is used in the corrector step. The alternate version employs a backward difference in the predictor step and a forward difference in the corrector step. For both versions, the predictor and corrector steps can be combined to yield

$$u_j^{n+1} = \frac{\delta}{8}\left(\frac{\delta}{4} + \tau\right)u_{j-2}^n + \frac{1}{2}\left(\frac{\delta}{2} + \tau + \tau^2 - \frac{\delta\tau}{2} - \frac{\delta^2}{4}\right)u_{j-1}^n + \left(1 - \frac{\delta}{2} + \frac{3}{16}\delta^2 - \tau^2\right)u_j^n + \frac{1}{2}\left(\frac{\delta}{2} - \tau + \tau^2 + \frac{\delta\tau}{2} - \frac{\delta^2}{4}\right)u_{j+1}^n + \frac{\delta}{8}\left(\frac{\delta}{4} - \tau\right)u_{j+2}^n \tag{D.3}$$

Let

$$A(\theta) \stackrel{def}{=} \left(1 - \frac{\delta}{2} + \frac{3}{16}\delta^2 - \tau^2\right) + \left(\frac{\delta}{2} + \tau^2 - \frac{\delta^2}{4}\right)\cos\theta - i\tau\left(1 - \frac{\delta}{2}\right)\sin\theta + \frac{\delta^2}{16}\cos 2\theta - \frac{i\delta\tau}{4}\sin 2\theta \tag{D.4}$$

and

$$\underline{u}_M(j,n,k) \stackrel{def}{=} b_k e^{ij\theta_k}[A(\theta_k)]^n \tag{D.5}$$

where b_k are defined in Eq. (5.8). Then the solution to the finite difference problem defined by Eqs. (D.3) and (B.2) is

$$u_j^n = \underline{u}_M(j,n) \stackrel{def}{=} \sum_{k=0}^{K-1} \underline{u}_M(j,n,k) \tag{D.6}$$

Combining Eqs. (5.17) and (D.4), it can be shown that

$$\varepsilon(\theta) \stackrel{def}{=} \frac{A(\theta)}{A_a(\theta)} - 1 = \frac{i\tau(1-\tau^2)}{6}\theta^3 + \frac{1}{48}[\delta(1-6\tau^2) - 6\tau^2(1-\tau^2)]\theta^4 + O(\theta^5) \tag{D.7}$$

By using the arguments used to establish Eq. (5.40), one can show that

$$[A(\theta)]^n - [A_a(\theta)]^n \doteq n[A_a(\theta)]^n \varepsilon(\theta) \qquad \text{if } n|\varepsilon(\theta)| \ll 1 \tag{D.8}$$

Eq. (5.65) follows directly from Eqs. (5.64), (D.5), (5.53), (D.8), and (D.7).

We conclude this appendix with a discussion on the operation count of the MacCormack scheme. Because the coefficients in front of the mesh variables on the right side of Eq. (D.3) are constants, they need not be reevaluated during the numerical marching. Thus, for each j, the MacCormack scheme requires 5 multiplications and 4 additions to advance one time step.

Appendix E

The proofs for several assertions made in Section 6 are provided here.

We begin with the assertion: Let $\mu > 0$. Then the Lax stability of the MacCormack scheme for solving Eq. (2.2) requires that the mesh be refined such that the parameter δ remains bounded as $\Delta t, \Delta x \to 0$.

Proof: Because *(i)* the sum of the terms involving δ^2 on the right side of Eq. (D.4) = $(\cos\theta - 1)^2 \delta^2 / 8$, and *(ii)* $\lim_{\Delta x \to 0} \tau/\delta = 0$, one concludes that, for any θ with $\cos\theta \neq 1$, the amplification factor $A(\theta)$ of the MacCormack scheme will become unbounded as $\Delta t, \Delta x \to 0$ if δ becomes unbounded as $\Delta t, \Delta x \to 0$. Since uniform-boundedness of the spectral radius of the amplification matrix is necessary for the Lax stability [p.70, 4], the proof is completed. Q.E.D.

Next we show that Eqs. (6.29) and (6.30) are true if $u = \tilde{u}(x,t)$ and $v = \tilde{v}(x,t)$ satisfy Eq. (6.27), i.e., $\tilde{w}(x,t) = 0$. To proceed, note that

$$1 > \left| \frac{1 - \tau^2 - \delta}{1 - \tau^2 + \delta} \right| \quad \text{and} \quad 1 > \frac{1 - \tau^2}{1 - \tau^2 + \delta} > 0 \tag{E.1}$$

follow directly from Eq. (6.18). Also,

$$\left[\tau \pm \frac{\delta(1 \mp \tau)}{1 - \tau^2 + \delta} \right] \frac{1 \pm \tau}{\Delta t} \times O[(\Delta x)^4] = \left[a(\Delta x \pm a \Delta t) \pm 4\mu \frac{1 - \tau^2}{1 - \tau^2 + \delta} \right] \times O[(\Delta x)^2] \tag{E.2}$$

$$\left[\tau \pm \frac{\delta(1 \mp \tau)}{1 - \tau^2 + \delta} \right] \frac{1 - \tau^2 - \delta}{\Delta t} \times O[(\Delta x)^4]$$

$$= \left\{ a[(\Delta x)^2 - a^2(\Delta t)^2 - 4\mu(\Delta t)] \pm 4\mu(\Delta x \mp a \Delta t) \frac{1 - \tau^2 - \delta}{1 - \tau^2 + \delta} \right\} \times O[(\Delta x)] \tag{E.3}$$

With the aid of Eqs. (6.23) and (E.1) - (E.3), and the fact that $\partial \tilde{w} / \partial x = 0$ if $\tilde{w}(x,t) = 0$, one concludes that Eq. (6.29) is true if $u = \tilde{u}(x,t)$ and $v = \tilde{v}(x,t)$ satisfy Eq. (6.27).

Next, we have

$$\tau \to 0 \quad \text{and} \quad \Delta x \, \delta \to 0 \qquad \text{as } \Delta t / \Delta x \to 0 \tag{E.4}$$

Also,

$$(1 - \tau^2 + \delta)^2 (\Delta x)^2 = (\Delta x - \tau^2 \Delta x + \delta \Delta x)^2 \tag{E.5}$$

$$\frac{(1 - \tau^2 + \delta)^2}{\delta} \times O[(\Delta t)^2] = \frac{(\Delta x - \tau^2 \Delta x + \delta \Delta x)^2}{4\mu} \times O[(\Delta t)] \tag{E.6}$$

$$\left[1 \pm \frac{\tau}{\delta}(1-\tau^2) \right] = 1 \pm \frac{a\,\Delta x}{4\mu}(1-\tau^2) \tag{E.7}$$

$$(1-\tau^2-\delta) \times O[(\Delta x)^2] = O[(\Delta x)^2] + O[(\Delta t)^2] - \delta\,\Delta x \times O[(\Delta x)] \tag{E.8}$$

Eq. (6.30) now follows directly from Eqs. (6.24) and (E.4) - (E.8).

If $\mu \neq 0$ and the rule of mesh refinement is such that δ remains bounded as Δt, $\Delta x \rightarrow 0$, then *(i)* $\Delta t = O[(\Delta x)^2]$ and *(ii)* $\Delta t / \Delta x \rightarrow 0$ as Δt, $\Delta x \rightarrow 0$. With the aid of Eqs. (E.1) - (E.8), it is easy to show that Eq. (6.31) is true if $u = \tilde{u}(x,t)$ and $v = \tilde{v}(x,t)$ satisfy Eq. (6.27) and the rule of mesh refinement is such that δ remains bounded as Δt, $\Delta x \rightarrow 0$.

REFERENCES

1. W. Kaplan, Advanced Calculus (Addison-Wesley, Reading, MA, 1952).

2. T.M. Apostol, Mathematical Analysis (Addison-Wesley, Reading, MA, 1957).

3. D.A. Anderson, J.C. Tannehill and R.H. Pletcher, Computational Fluid Mechanics and Heat Transfer (McGraw-Hill, New York, NY, 1984).

4. R.D. Richtmyer and K.W. Morton, Difference Methods for Initial-value Problems, 2nd edition (Interscience, 1967).

5. G. Strang, Linear Algebra and Its Applications, 2nd edition (Academic Press, Orlando, FL, 1980).

6. B. Noble and J.W. Daniel, Applied Linear Algebra, 2nd edition (Prentice-Hall, Englewood Cliffs, NJ, 1977).

7. L. Bers, F. John and M. Schechter, Partial Differential Equations (Interscience, 1964).

	a	μ	b	K	t	Re	τ_0	τ_*	$\overline{\overline{\tau}}_0$
# 1	1.0	0.1	0.	30	0.5	1/12	0.048002	0.048011	0.1064
# 2	1.0	0.1	0.	30	1.0	1/12	0.048002	0.048026	0.1064
# 3	1.0	0.1	0.	60	0.5	1/24	0.024042	0.024043	0.05364
# 4	1.0	0.01	0.	30	3.1	5/6	0.40299	0.40309	0.6380

Table 7.1. Definitions of the problem sets #1 - #4 and the corresponding values of Re, τ_0, τ_*, and $\overline{\overline{\tau}}_0$.

	a	μ	K	n	Δt	δ	τ_+	τ_*	$\overline{\overline{\tau}}_+$
# 5	1.0	0.01	30	390	$1/(45\sqrt{3})$	$0.8/\sqrt{3}$	0.4472	0.4468	0.8014
# 6	1.0	0.01	30	312	$1/(36\sqrt{3})$	$1/\sqrt{3}$	0.	0.	0.7435

Table 7.2. Definitions of the problem sets #5 and #6, and the corresponding values of δ, τ_+, τ_*, and $\overline{\overline{\tau}}_+$.

	a	μ	b	K	t
# 7	0.	0.01	0.	30	15.
# 8	1.0	0.	0.	30	0.5

Table 7.3. Definitions of problem sets #7 and #8.

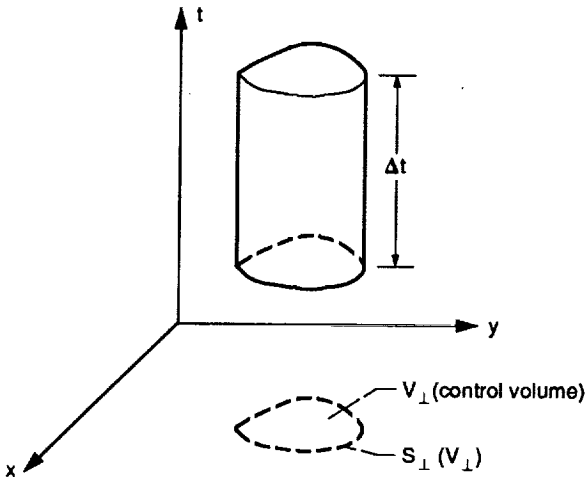
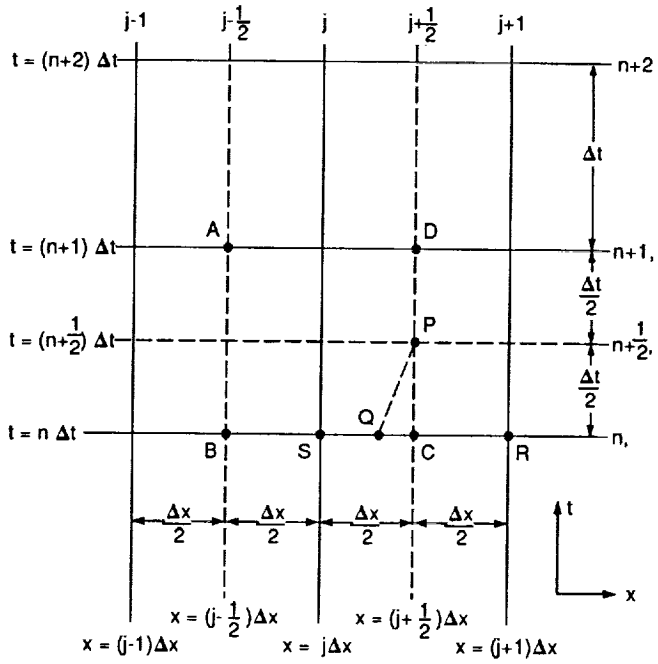
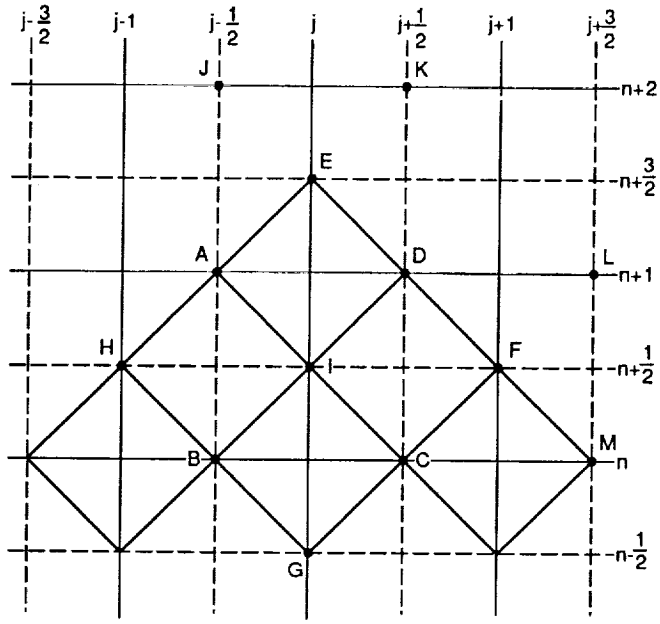


Figure 1.1—A cylinder in a space-time E_3.

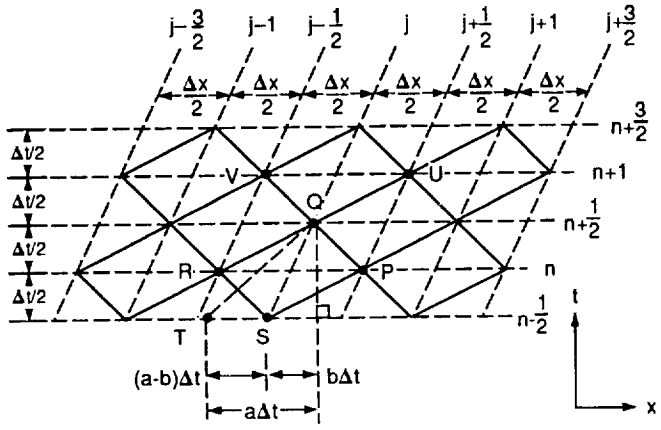


(a) A uniform mesh on the x-t plane (n = 0, 1, 2, ...; j = 0, ±1, ±2, ...).

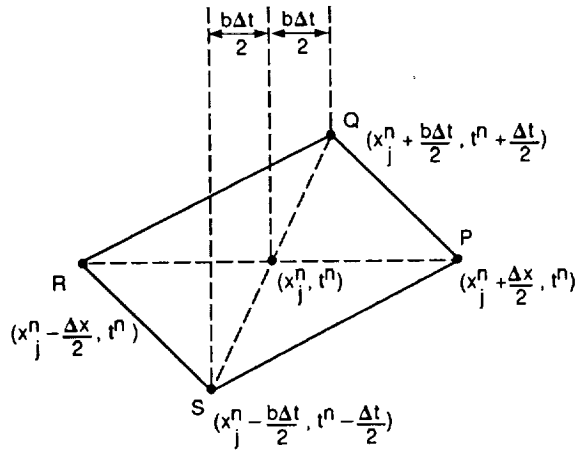


(b) The solution elements and conservation elements of the Lax-Wendroff scheme.

Figure 1.2.

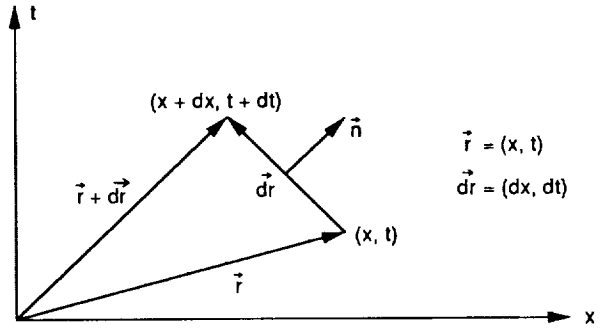


(a) A uniform moving mesh in a space-time E_2.

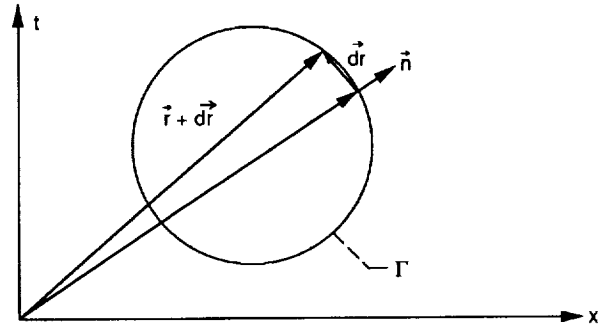


(b) A conservation element of the present scheme.

Figure 2.1.



(a) A line segment in a space-time E_2.



(b) A simple closed curve Γ in a space-time E_2.

Figure 2.2.

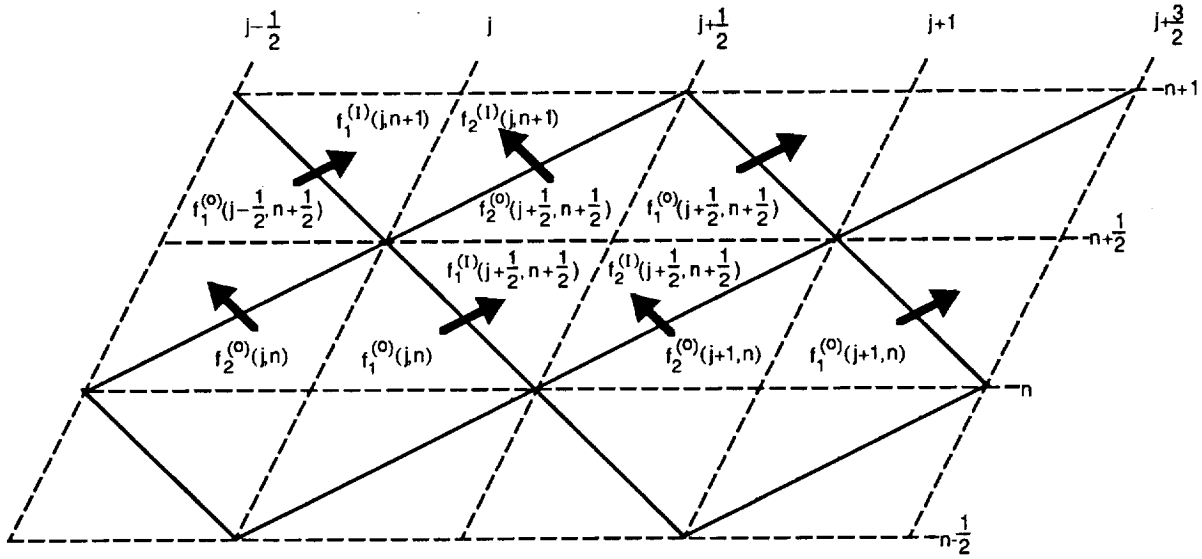


Figure 2.3.—Interface flux conservation relation.

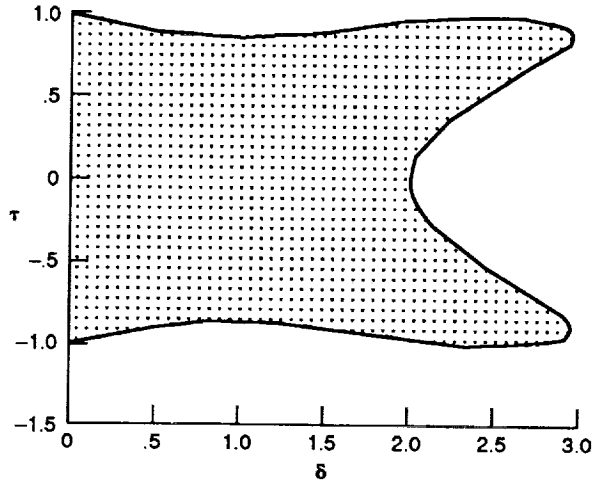


Figure 4.1—Stability region (shaded area) of the MacCormack scheme on the δ-τ plane.

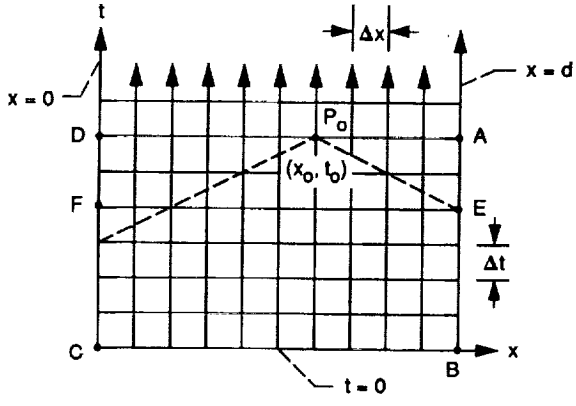


Figure 6.1.—Computation domain with d ≥ x ≥ 0 and t ≥ 0.

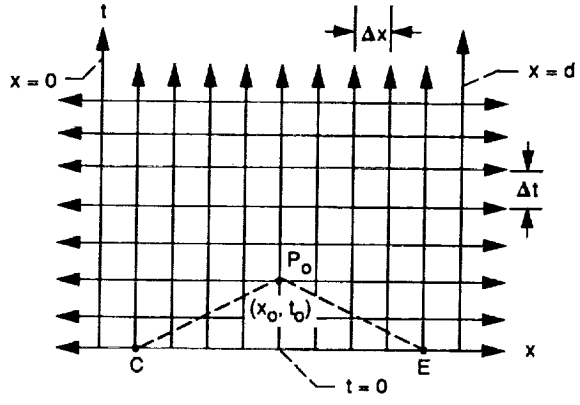


Figure 6.2.—Computation domain with + ∞> x > − ∞ and t ≥ 0.

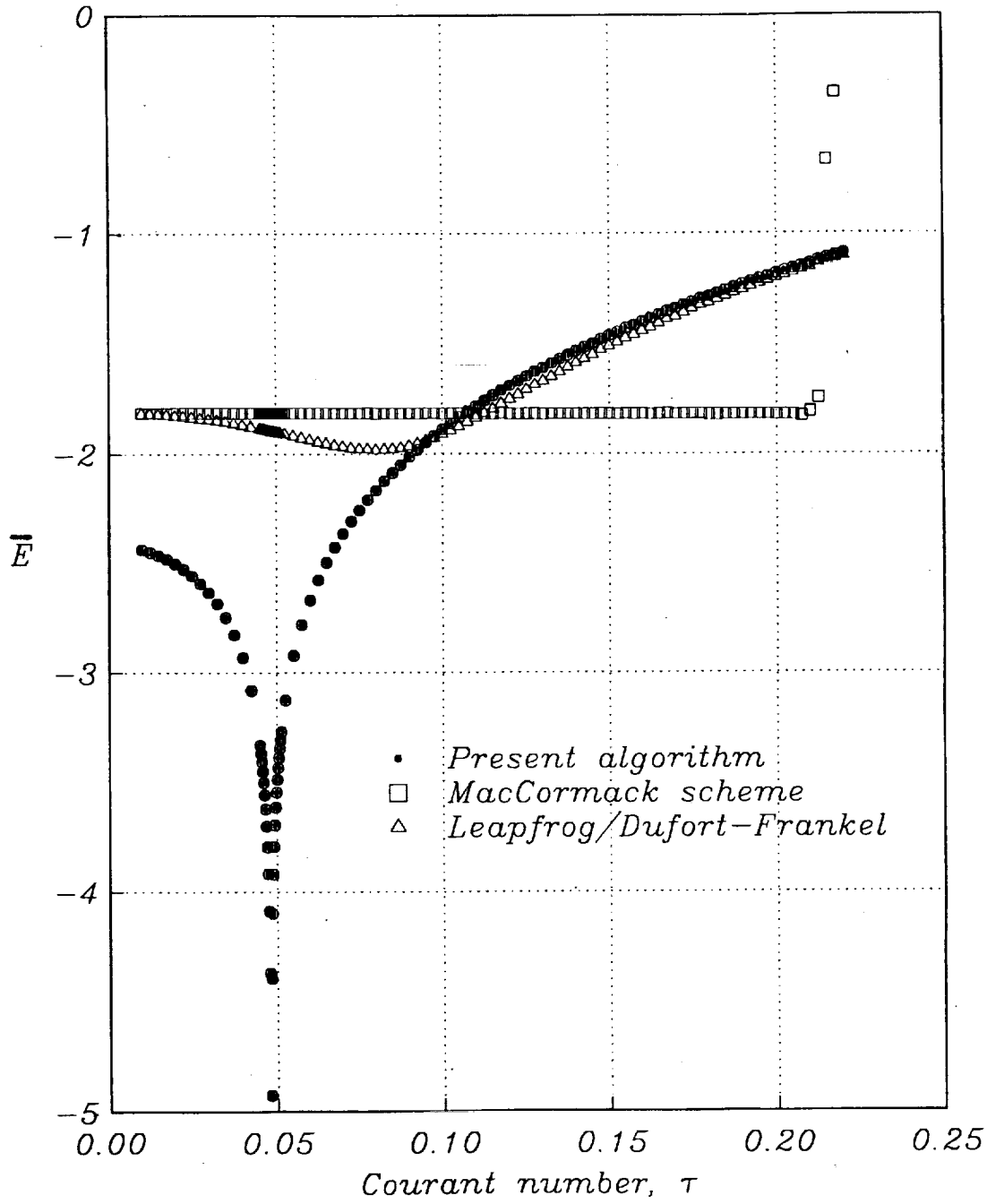


Figure 7.1—Accuracy of test problems in Set #1 ($\tau = 30\Delta t$).

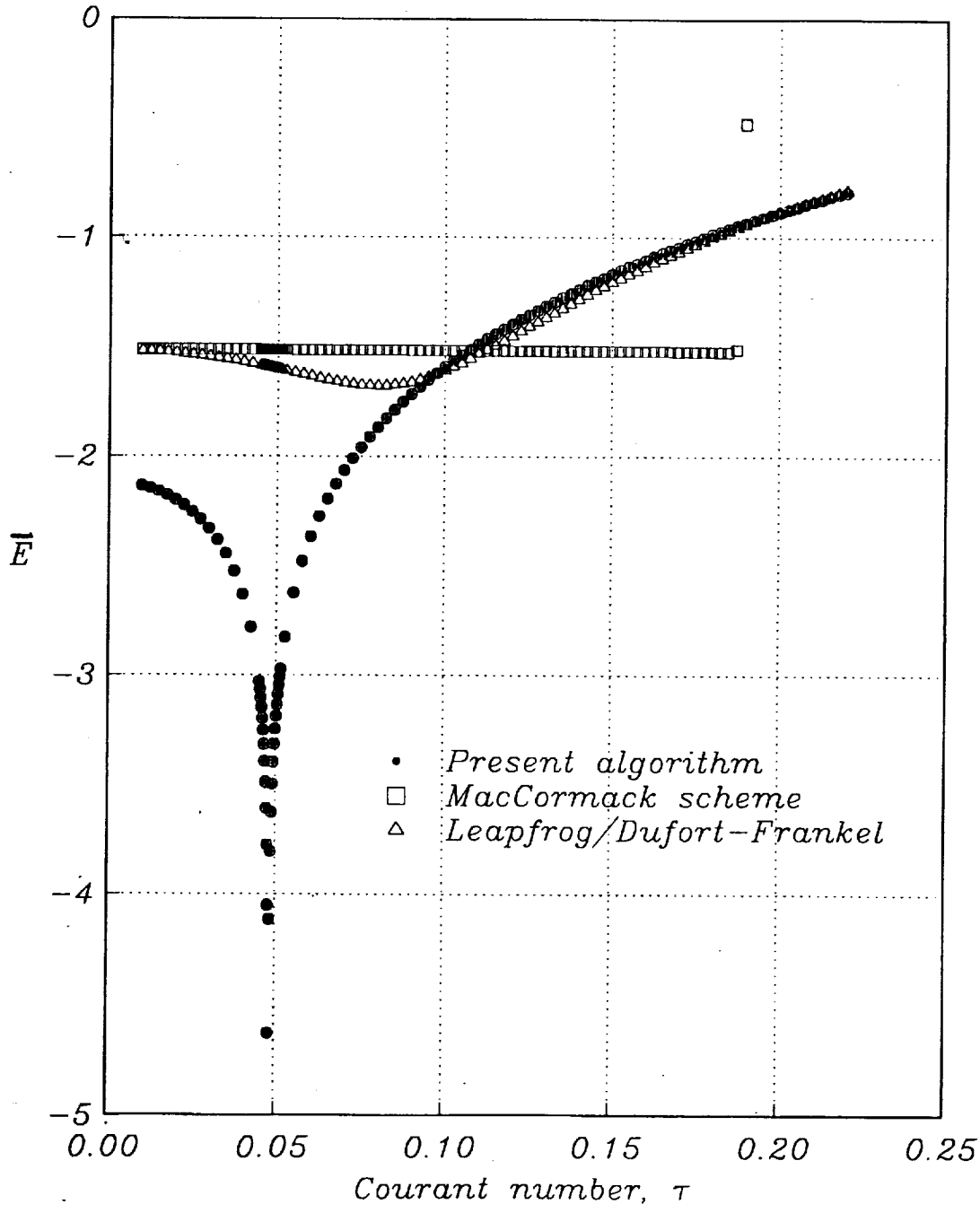


Figure 7.2—Accuracy of test problems in Set #2 ($\tau = 30\Delta t$).

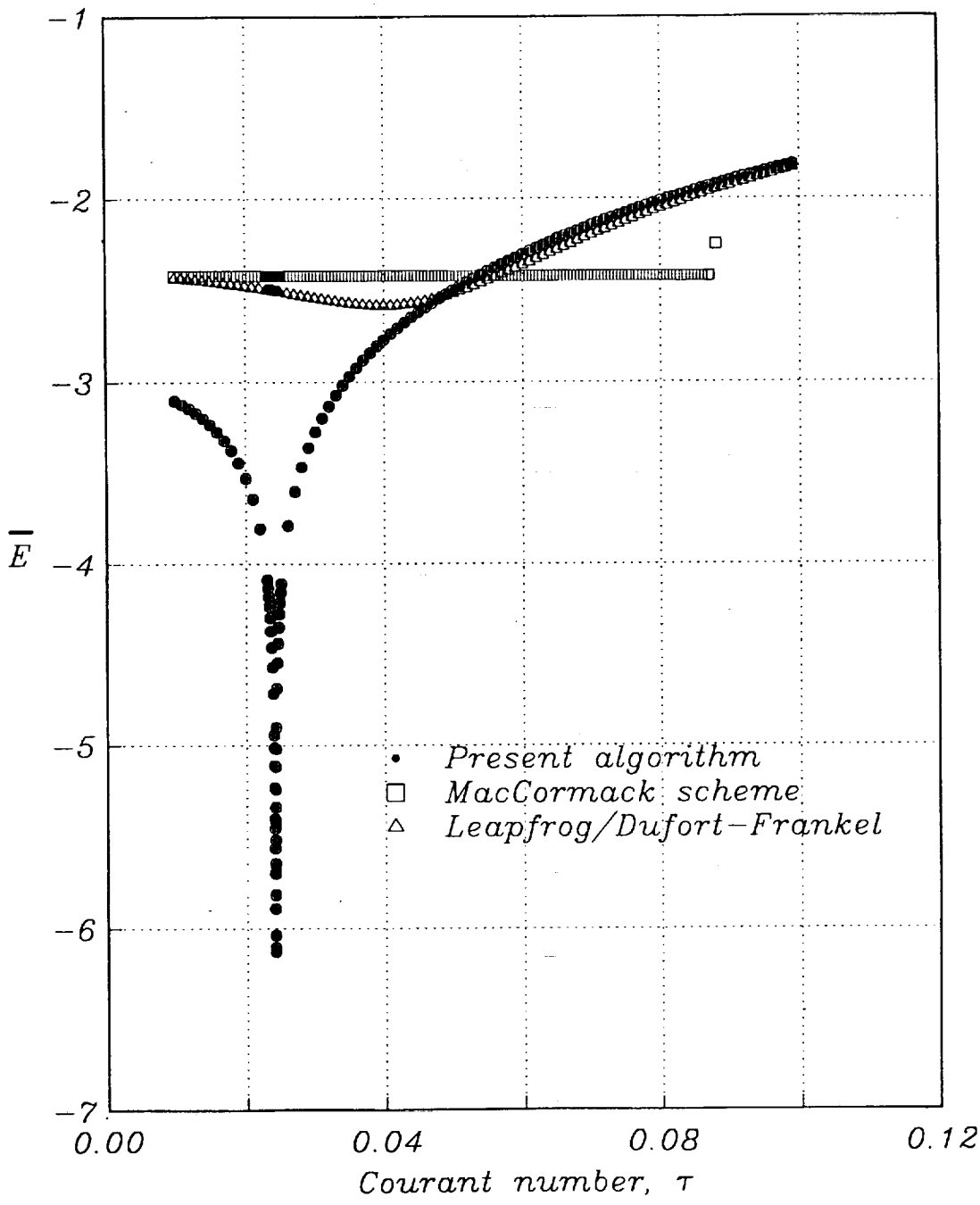


Figure 7.3—Accuracy of test problems in Set #3 ($\tau = 60\Delta t$).

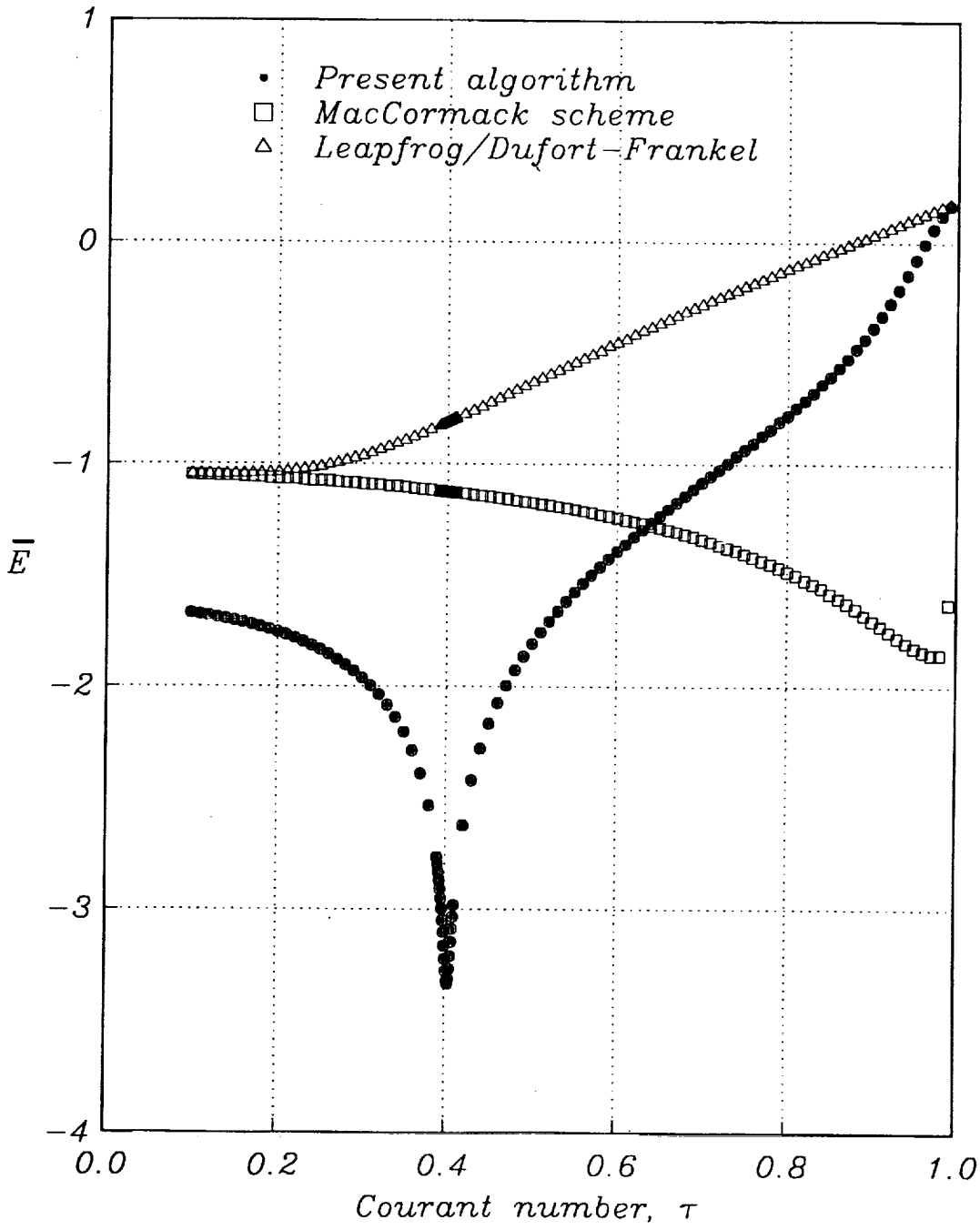


Figure 7.4—Accuracy of test problems in Set #4 ($\tau = 30\Delta t$).

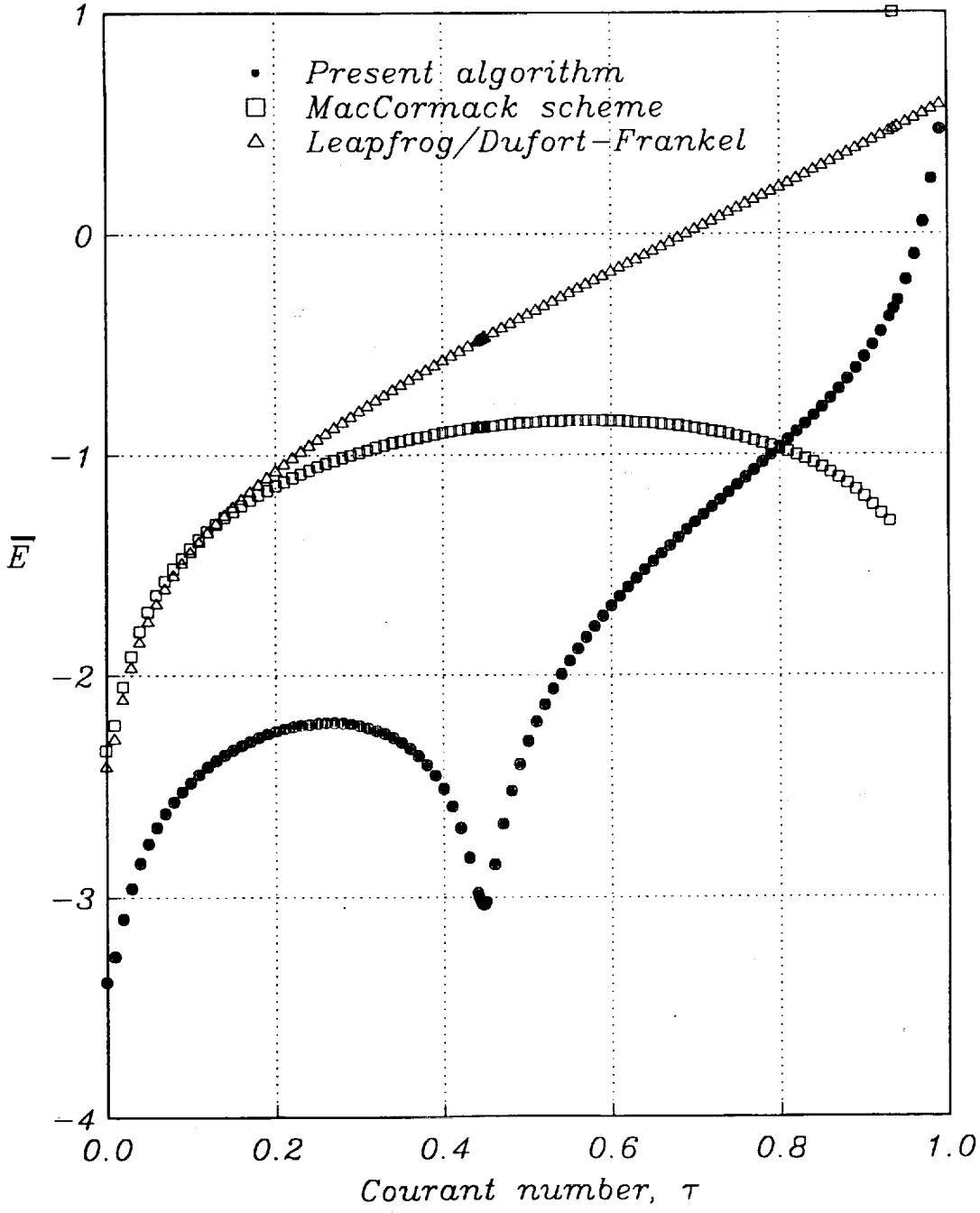


Figure 7.5—Accuracy of test problems in Set #5 ($\tau = 2(1-b)/(3\sqrt{3})$).

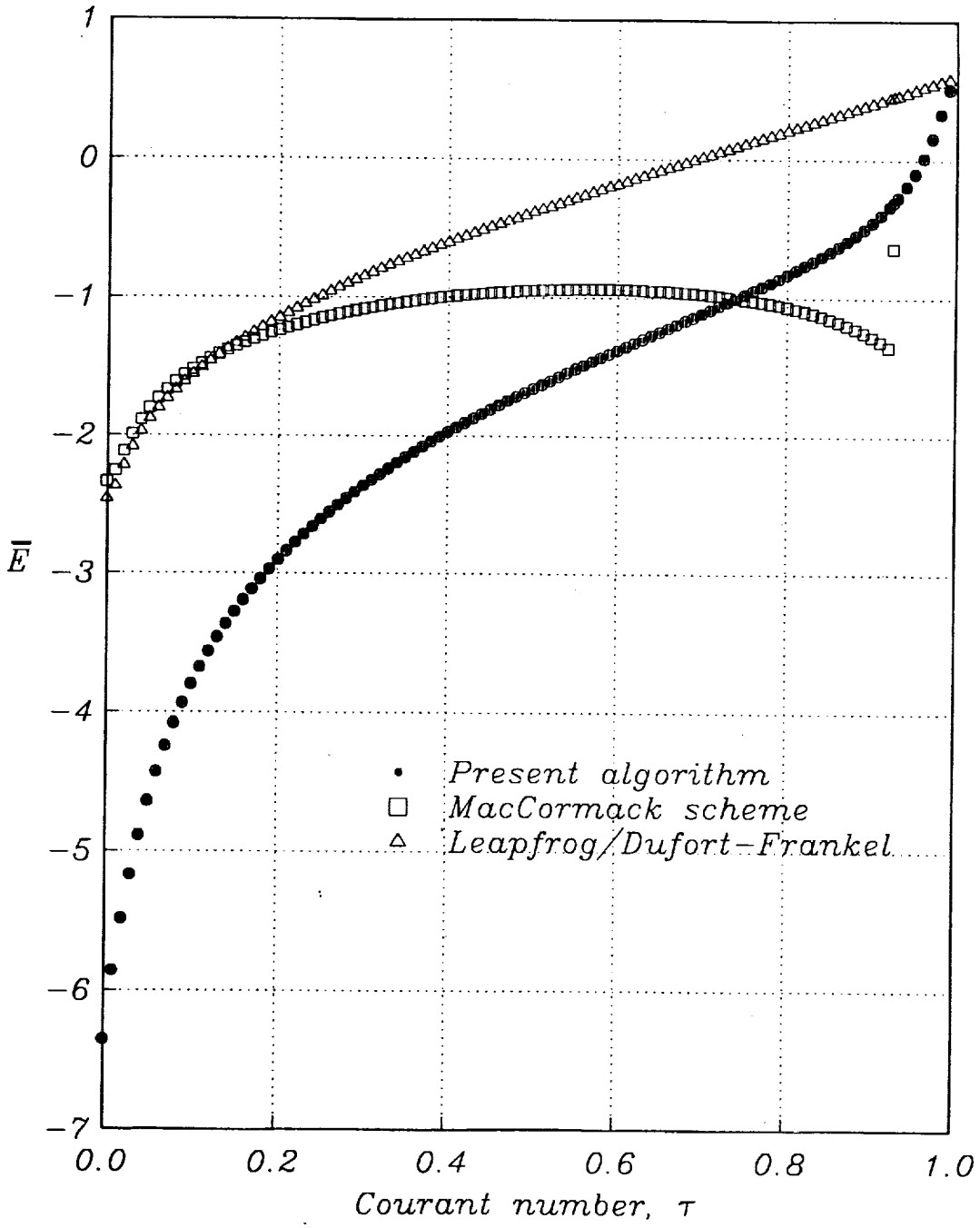


Figure 7.6—Accuracy of test problems in Set #6 ($\tau = 5(1-b)/(6\sqrt{3})$).

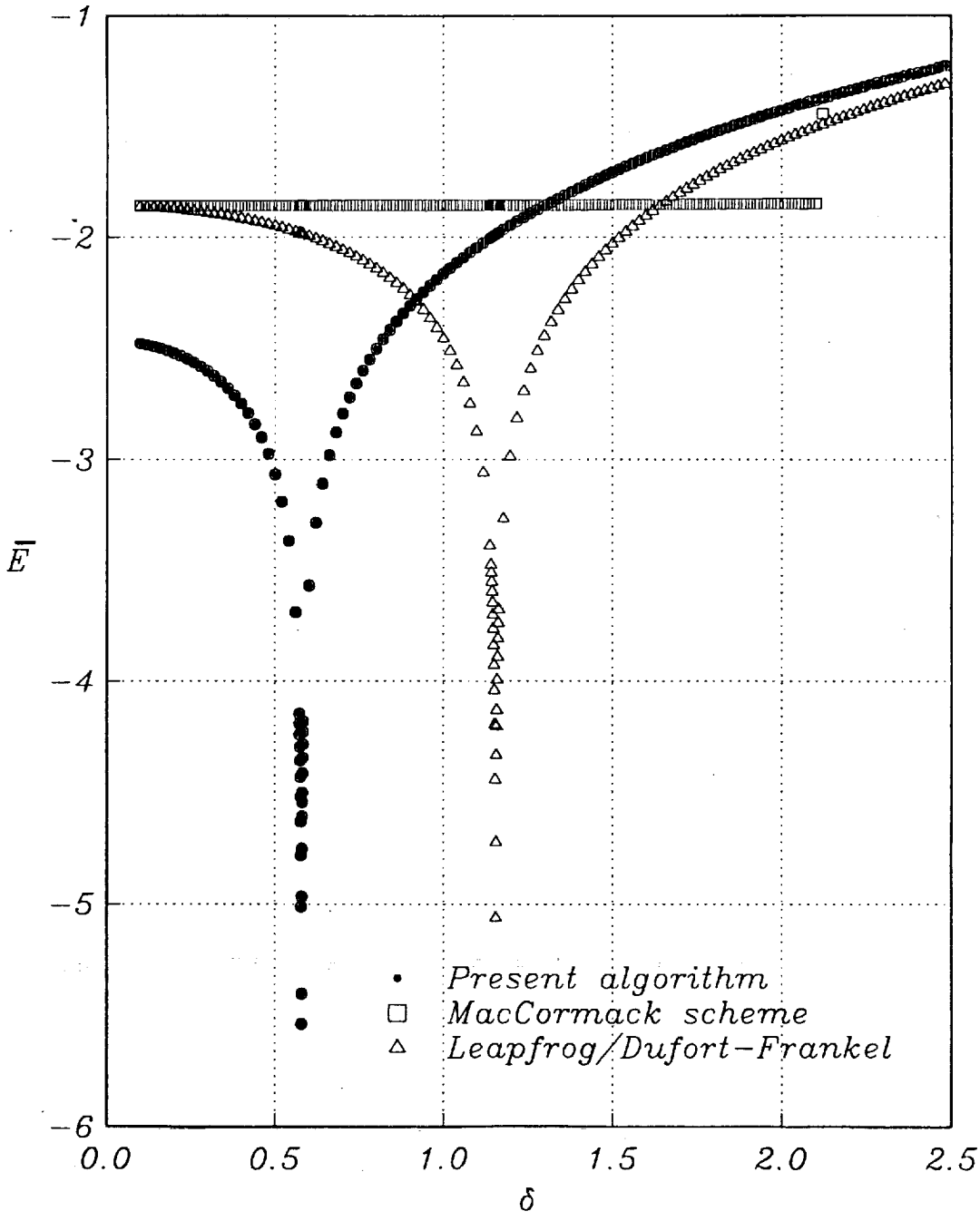


Figure 7.7—Accuracy of test problems in Set #7 ($\delta = 36\Delta t$).

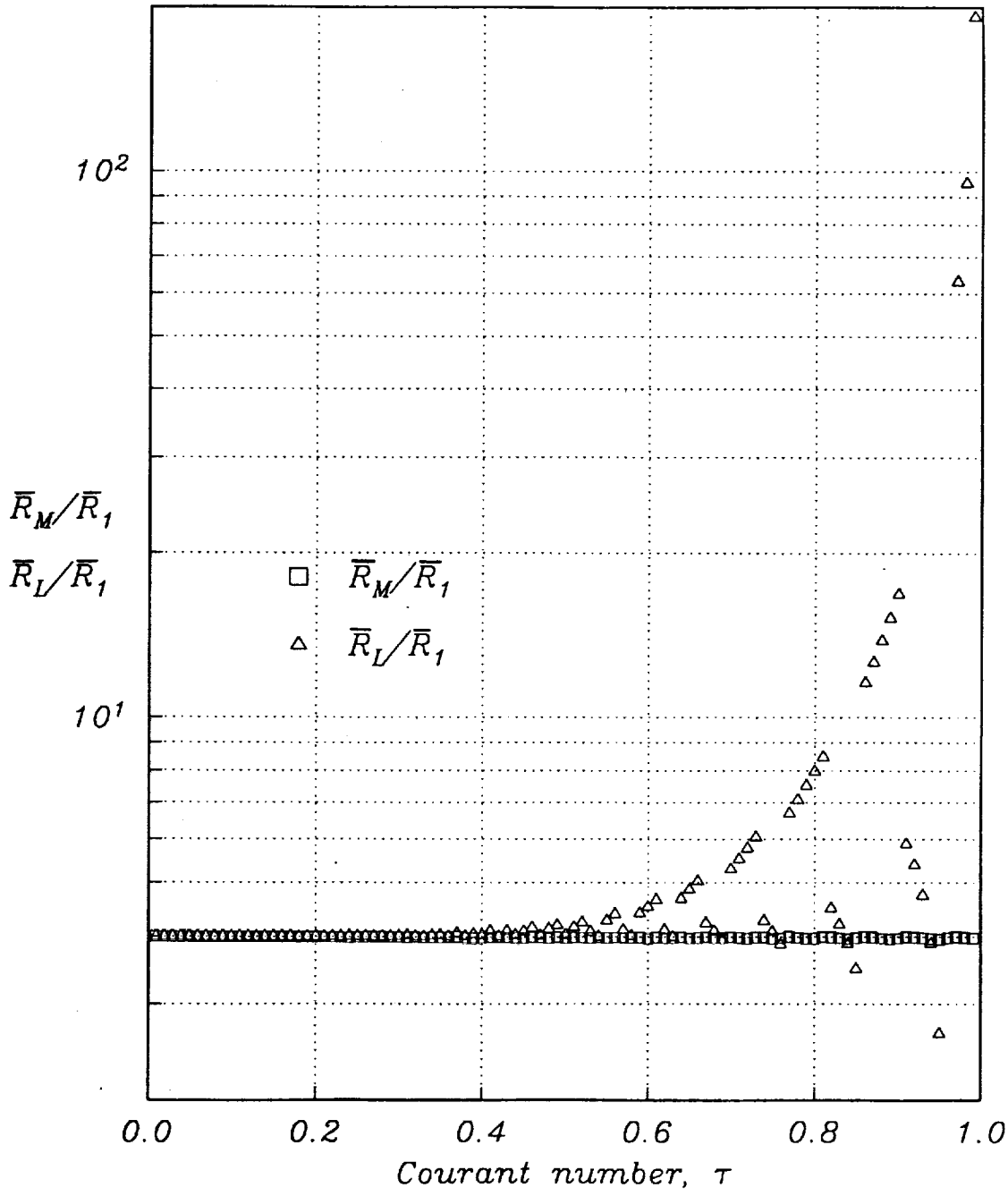


Figure 7.8—Accuracy of test problems in Set #8 ($\tau = 30\Delta t$).

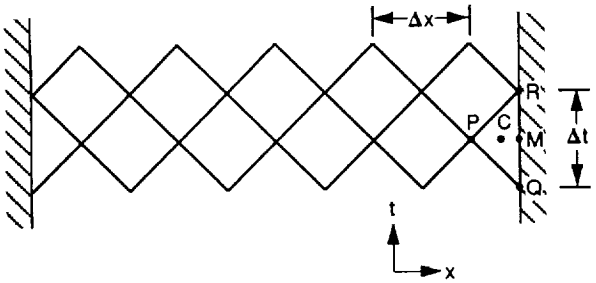


Figure 8.1.—Boundary conservation element (ΔPQR).

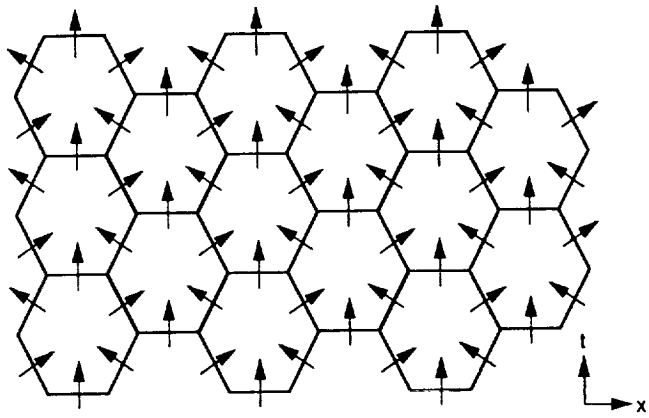


Figure 8.2.—Hexagons as conservation elements.

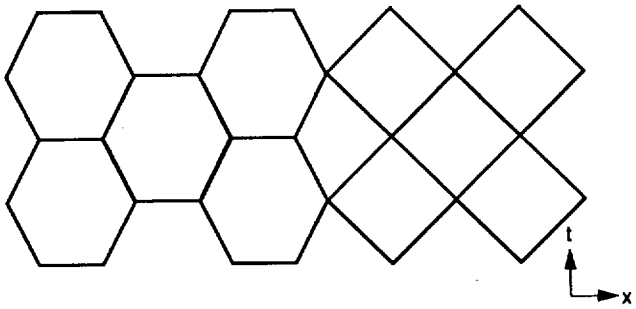


Figure 8.3.—Space-time E_2 divided into conservation elements of different geometric shapes.

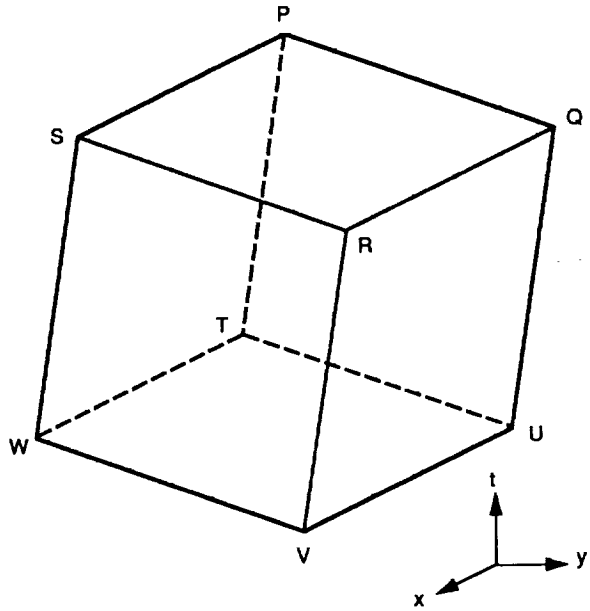


Figure 8.4.—Conservation element in space-time E_3.

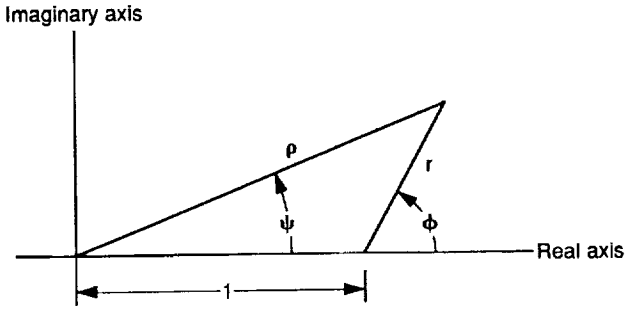


Figure C.1.—Geometric relations among the parameters r, ϕ, ρ and ψ defined by Eqs. (C.11) and (C.12).

NASA
National Aeronautics and Space Administration

Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA TM - 104495		

4. Title and Subtitle
A New Numerical Framework for Solving Conservation Laws-The Method of Space-Time Conservation Element and Solution Element

5. Report Date
August 1991

6. Performing Organization Code

7. Author(s)
Sin-Chung Chang and Wai-Ming To

8. Performing Organization Report No.
E - 6403

10. Work Unit No.
505 - 62 - 52

9. Performing Organization Name and Address
National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135 - 3191

11. Contract or Grant No.

13. Type of Report and Period Covered
Technical Memorandum

12. Sponsoring Agency Name and Address
National Aeronautics and Space Administration
Washington, D.C. 20546 - 0001

14. Sponsoring Agency Code

15. Supplementary Notes
Sin-Chung Chang, NASA Lewis Research Center; Wai-Ming To, Sverdrup Technology, Inc., Lewis Research Center Group, 2001 Aerospace Parkway, Brook Park, Ohio 44142. Responsible person, Sin-Chung Chang, (216) 433 - 5874.

16. Abstract

A new numerical framework for solving conservation laws is being developed. This new approach differs substantially from the well established methods, i.e., finite difference, finite volume, finite element, and spectral methods, in both concept and methodology. It employs (1) a nontraditional formulation of the conservation laws in which space and time are unified and treated on the same footing and (2) a nontraditional use of discrete variables such that numerical marching can be carried out by using a set of relations that represents both local and global flux conservation. In this paper, we also (1) explore the concept of a dynamic space-time mesh and the need for a unified treatment of physical variables and mesh parameters; (2) study the stability, dissipation and dispersion of the current scheme by using a rigorous Fourier analysis; (3) develop a new error analysis technique that enables us to predict and interpret the numerical errors of the current and other classical schemes; (4) study the consistency and truncation error of the current scheme; and (5) compare the errors of the numerical solutions generated by the current scheme and other classical schemes.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Space-time Conservation element Solution element	Unclassified - Unlimited Subject Category 64

19. Security Classif. (of the report)	20. Security Classif. (of this page)	21. No. of pages	22. Price*
Unclassified	Unclassified	-114-	A04

NASA FORM 1626 OCT 86

*For sale by the National Technical Information Service, Springfield, Virginia 22161